SRTTU

Journal of Electrical and Computer Engineering Innovations

JECEI, Vol. 6, No. 1, 2018

Regular Paper

A New Model for Text Coherence Evaluation Using Statistical Characteristics

Mohamad Abdolahi[1,*] and Morteza Zahedi[1]

[1]Kharazmi International Campus, Shahrood University, Shahrood, Iran.

[*]Corresponding Author's Information: mabdolahi512@yahoo.com

ARTICLE INFO

ARTICLE HISTORY:
Received 06 February 2018
Revised 29 March 2018
Accepted 07 April 2018

KEYWORDS:
Local text coherence
Global text coherence
Word vector
Word embeddings
Word2vec algorithm

ABSTRACT

Discourse coherence modeling evaluation becomes a critical but challenging task for all content analysis tasks in Natural Language Processing subfields, such as text summarization, question answering, text generation and machine translation. Existing methods like entity-based and graph-based models are engaging in semantic and linguistic concepts of a text. It means that the problem cannot be solved very well and these methods are only very limited to available word co-occurrence information in the sequential sentences within a short part of a text. One of the greatest challenges of the above methods is their limitation in long documents coherence evaluation and being suitable for documents with low number of sentences. Our proposed method focuses on both local and global coherence. It can also assess the local topic integrity of text at the paragraph level regardless of word meaning and handcrafted rules. The global coherence in the proposed method is evaluated by sequence paragraph dependency. According to the derived results in word embeddings, by applying statistical approaches, the presented method incorporates the external word correlation knowledge into short and long stories to assess both local and global coherence, simultaneously. Using the effect of combined word2vec vectors and most likely n-grams, we show that our proposed method is independent of the language and its semantic concepts. The derived results indicate that the proposed method offers the higher accuracy with respect to the other algorithms, in long documents with a high number of sentences.

1. INTRODUCTION

Coherency, as a property of well-written texts, helps us to read and understand them easier than a random sequence sentences. Although the same information can be organized in multiple ways to create a coherent text, some forms of text organization will be indisputably judged incoherent. In recent years, there have been several investigations into text coherence evaluation. There are also high quality systems that are designed with the ability to produce texts very close to those written by human. Using automatic methods for evaluating or increasing the quality of coherence is considered the most important goal of all text processing systems such as statistical machine translation [1]-[4], text generation, mode detection, question answering, student essay scoring [5]-[7] and text summarization [8]-[10]. There is a growing body of literature that recognizes the importance of integrity of text processing output. There exist increasing attentions to some non-statistical approaches to assess the coherence of text processing outputs. However, major problems with them are semantic difficulties and a complete understanding of linguistic concepts.

Two main categories for text coherence evaluation are local and global. Local coherence is the well connectedness of adjacent sentences through lexical

cohesion [11] or entity repetition whereas the global one is the discourse-level relation connecting remote sentences or adjacent paragraphs [12]. The main challenge faced by many experiments is that none of the previous methods can evaluate the coherence on both levels [13]. Methods such as entity-based, graph-based and approaches have been known as entropy-based methods are the well-known methods that had been proposed and used in most articles.

While some research has been carried out on local coherence evaluation, there is still very little scientific understanding of both local and global coherence, simultaneously. The coherence model we present, fall into the popular and efficient method of Google "Word2vec". Our proposed method focuses on both local and global coherence, which assess the local topic integrity of text at the paragraph level regardless of word meaning and handcrafted rules and global coherence is evaluated by sequence paragraph dependency.

Previously published studies are limited to global coherence and there is a semantic relatedness between all sentences and title or topic subject of the document. Some studies have shown the beneficial effects of entity-based approaches to evaluate sentences dependency of documents, but our method showed a deterioration of them in long documents with more sentences. The purpose of this investigation is to explore the two novel advantages: Firstly, usually a paragraph is a big part of each document and the subject integrity of each paragraph as a local cohesive unit is previously assessed. Secondly, the number of paragraphs in a text is much less than the number of its sentences, hence, evaluating the subject dependency of few paragraphs is very simple operation than all sentences dependency in the document. Both qualitative and quantitative methods that consist of word embeddings, n-grams, word2vec vectors and numerical matrices of sentences were used in this investigation.

The overall structure of the study takes the form of nine sections. This paper begins by introduction method and it will then go on to give a brief overview of the recent history of text coherence evaluation. The third section is concerned with the text preprocessing methods used for this study. The fourth and fifth sections present a brief introduction on the state of word embeddings and word2vec Google algorithm. Section six begins by laying out the proposed method and theoretical dimensions of the research, and Section seven looks at our data set. Section eight analyses the results of interviews and focus on group discussions undertaken during evaluation method and finally conclusion is drawn in Section nine.

2. Related Works

In this section, we briefly describe the related previously proposed methods. The first systematic study of text coherence evaluation was reported by Foltz et al. in 1998 [14]. According to this study, text coherence is a function of semantic relatedness between two adjacent sentences within a text. Then, a vector-based representation of lexical meaning is used to compute the semantic relatedness between sequence sentences. So far, several supervised approaches have been identified as being potentially important: entity-based model [15]-[20], discourse relation-based model [21], syntactic patterns-based model [22], co-reference resolution-based model [23], [24], content-based model via Hidden Markov Model [25], [26] and cohesion-driven based model [27]. These methods compute the relationship topics in adjacent sentences to obtain the coherence in a supervised way.

A. Entity-Based Model

Entity-based accounts of local coherence have been popular within the linguistic and cognitive science. It is also one of the most famous approaches that analyses the grammatical role of words in adjacent sentences, to extract patterns from them and assess local coherence [28]. This model was proposed by R. Barzilay, M. Lapata for the first time [15], [16], [29], but some other combined novel approaches such as neural network models [30] and original bipartite graph [31] was proposed in recent years. Essay scoring is other scope that uses entity-based method. J. Burstein combined entity-based features with aspect related to grammar errors and words usage to improve the performance of automated coherence prediction for student essays [6].

B. Graph-Based Model

Historically, research investigating the factors associated with graph theory has focused on many NLP tasks. Some drawbacks and shortcomings of entity-based model and much of graph theory benefits, made some researches to focus on identifying and evaluating novels with combined graph and entity-based methods. Strube and Guinaudeau proposed an approach that offers a combination of entity grade and graph-based model to overcome the limitation ability of entity grade to detect consistency in just neighboring sentences [32]. Petersen and Simonsen proposed a model which is a combination of graph theory and entropy method for assessing the consistency of document sentences [19]. In theis model, by increasing more nouns in the document, more peripheral information is participating in the context which led to lower the global coherence. Other graph-based coherence features were introduced by M. Mesgar [33], which

are based on frequent subgraphs. Herein, the coherence texts are consistent of particular patterns in their extracted subgraphs.

C. Statistical Machine Translation Algorithms

Some systematic reviews of EM and IBM algorithms in statistical machine translation have been undertaken in text coherence evaluation [2], [34]. The main idea of statistical machine translation is the meaning of each word in the target language introduced several words. Therefore, each word lead to link into multiple sentences and the algorithm chooses the most likely sentences.

D. Lexical Chain Models

Over the past decade, most research in text coherence evaluation has emphasized the use of Lexical chains approaches. They provide a representation of the lexical cohesion structure of a text such that the words of a text can be presented by features introduced in the previous section and causes the conceptual and thematic relationship between sentences in a document. Early examples of research into lexical chain include D. Xiong proposed method for evaluation of machine translation output [2]. S. Somasundaran et al. focused on lexical chaining methods for measuring discourse coherence quality in test-taker essays [17].

E. Neural Network Models

Limitation of semantic features forces us to use modern approaches. These approaches try to extract the syntactic representation of discourse coherence by neural network approaches [35]. Long Short Term Memory networks (LSTM) have been used in the assessment of coherent texts [36]. The approach introduced by J. Li and D. Jurafsky offered two distinct models that manufacturers used to assess cohesion. L. Logeswaran and H. Lee mentioned the method based on neural approach to do the sentence ordering problem. This novel approach tried to use RNNs for sequence modeling tasks [38]. Deep neural network is the most recent model that tries to assess local and global text coherence [35].

3. TEXT PREPROCESSING

There are many structural variants for the words appeared in documents. So, it is needed to prepare input text for any text processing approach. It means that language and text processing algorithm, and different text preprocessing algorithm depends on the input text type. The major challenge is their restriction to a particular field and having no ability to apply and extend to all areas and languages. The most important preprocessing methods are tokenization, stop word removal, stemming and POS tagging. However, choosing the right text preprocessing method and the exploiting rate to text has a huge impact on the performance of the final processing algorithm, in the view of accuracy and speed. In the proposed method, we need all sentence components and applied preprocessing model is the same as that used in [39].

4. WORD EMBEDDINGS

Words often can map to vectors in a vector-space. The mapping is called embedding and it is used in many NLP tasks. The vectors are intended to reflect the usage, semantic similarities and relatedness of the words in the text that they represent. In other words, word embeddings reflect the meaning of the words relative to other words in the whole corpus. Firth (1957) introduced the powerful idea that the complete meaning of each word is related to its neighbors. This is one of the most successful idea of modern statistical natural language processing and after that it has been used very extensively in all NLP approaches. Before word embedding, most text processing algorithms often used the same general and existing methods that were introduced in the field of image processing and speech recognition [40], [42]. Word embeddings can capture subtle semantic relationships between words, such as the following well-known examples where vec (x) denotes the vector of word x [39].

vec (Berlin) –vec (Germany) + vec (France) = vec (Pairs) and vec (Einstein) – vec (scientist) + vec (Picasso) = vec (painter)

Despite the agreement of all methods on the ability of deep learning in text processing, a text has special features in comparison with other data like speech and image. One of the most important problems in the field of speech and image processing is the noise discovery, reduction or elimination. While, the most important text processing limitation is information lost and semantic ambiguities. In addition, in image processing, the processing system relies heavily on the information contained in the image itself and requires less background knowledge or external information. While in text processing external information and knowledge background can help much more to identify some of the existing ambiguities.

So far, there were many approaches using word embedding for text processing. Convolution neural network techniques are often presented as the best tool in most methods. Johnson's argument relies too heavily on qualitative analysis of convolution neural network to classify text [41], [42]. Nguyen and Grishman's interpretation overlooks text created filtering matrices and coding the distance between existing relationships of words and the initial training of algorithm [43]. Convolution neural network is also

applied on other text processing fields such as text summarization, question answering systems and text topic recognition [44]-[46]. Word embedding can be used in Completely Automated Public Turing test to tell Computers and Humans Apart methods (CAPTCHA) to make some smart CAPTCHA approaches [47].

5. WORD2VEC ALGORITHM

In the most of rule-based and statistical language processing algorithms words are assumed as atomic units in text.

However, all these previously mentioned methods suffer from some serious disadvantages such as sparse matrix with values of 1 and 0. In these methods values correspond to 1 are assumed as sentence word position and other positions are assumed 0. One major drawback of this approach is that generated matrices have no comprehensive information for considering word similarities or differences. Mikolov et al. introduced a novel word-embedding procedure, namely word2vec. This model learns a vector representation for each word using shallow neural network architecture. In order to predict nearby words, this neural network consists of an input layer, a projection layer, and an output layer [46].

This algorithm is able to guesses an acceptable word's meaning based on past appearances. One advantage of the word2vec is to gain insights into similar words being together in vector space. Word2vec applies a standard technique such as skip-gram on a given corpus.

The model avoids non-linear transformations and makes training extremely efficient. This enables learning of embedded word vectors from huge data sets with billions of words. Each word vector is trained to maximize the log probability of neighboring words in a corpus, given a sequence of words w_1 to w_T.

$$\frac{1}{T}\sum_{i=1}^{T}\sum_{j\in nb(t)}\log p(w_i|w_t) \tag{1}$$

where nb (t) is a set of neighboring words of word w_t and p (w_j | w_t) is the hierarchical softmax of the associated word vectors v_{wj} and v_{wt} .

Pennington et al. later introduced a new and different form of global log-bilinear regression model of word embedding; Glob2Vec that only utilizes local context windows [49]. Glob2Vec combines global word to word co-occurrence counts from a corpus, and local context windows based learning similar to word2vec to deliver an improved word vector representation.

6. THE PROPOSED APPROACH

To represent a text in word-level, each word can be represented as a numeric vector that is named word embedding. More specifically, the word is represented using a specific vector in the form of ew= $\{e_w^1, e_w^2, \ldots, e_w^K\}$, where K denotes the dimension of the word embedding. A text is coherent when there is a correlation between its components. In other words, a coherence document is a text made up of relatedness sentences. For the purpose of height measurement, sentences are considered as a smallest unit of a coherent text. In order to assess the integrity of paragraph, topic correlation between the sentences is evaluated. To assess the topic dependency and sentence relationship, sentences matrices are formed according to word2vec vector. The local coherence in our research is the sentences dependency in a paragraph. The first step in this process is to evaluate sentence dependency of each paragraph as local coherence and in the second step, paragraph dependency is assumed as global coherence. It means that the next sentence has a correlation to the (n-1) previous sentences according to the conditional probabilities.

$$\begin{aligned} p(t) &= p(S_1)p(S_2,|\,S_1)p(S_3|\,S_1,S_2)\ldots p(S_n|\,S_1\cdots S_{n-1}) \\ &= \prod_{i=1}^{n} p(S_n|\,S_1\cdots S_{n-i}) \end{aligned} \tag{2}$$

The estimation of $P(S_i \mid S_{i-1})$ is the comparison of sentence word2vec matrix features.

$$\begin{aligned} &p(S_i|\,S_{i-1}) = \\ &p((a_{(i-1)}\cdots a_{i-n})|\,a_{(i-1,1)},a_{(i-1,2)}),\ldots,a_{(i-1,m)}) \end{aligned} \tag{3}$$

where ($a_{(i-1)}$, $a_{(i-2)}$, … $a_{(i-n)}$) are features relevant for sentence S_i and $a_{(i-1,1)}$, $a_{(i-1,2)}$, …, $a_{(i-1,m)}$ for sentence S_{i-1}. The qualitative case studies we used, is a well-established approach in LD bigrams as proposed by R. Rosenfeld [37] and as employed in our own previous method [38]. The LD bigrams method is one of the more practical way of semantic integration of all text components text, especially for consecutive sentences. The semi-structured approach was chosen because in consecutive sentences, coherence patterns have consistent reducing pattern from one to five sentences. In order to understand this, sentences dependencies of more than five distances are almost constant. If the loss of coherence is calculated in five consecutive sentences in a paragraph, values have almost no significant changes. Given this idea, the best value for (n) is five.

A. Coherence vector

In this section, we discuss our method, Embedding-based Coherence Evaluation Model (ECEM), for

identifying and evaluating the topics relatedness details of a document. Our ECEM model includes three steps. First, we separate the document into its sentences. Then the normalized matrixes are built by distributed word embedding [49], [50]. At the end the sentence relatedness is evaluated based on their matrices dependencies. To assess the coherence of two sentences, we obtain two properties of their matrix similarity and inverse distance. To evaluate the similarity of the two sentences, the Cosine Similarity criterion (CS) of their matrix and to assess the inverse distance of them, the Manhattan Distance criterions (IMD) are obtained.

$$CS = \sum_{i=1}^{n} A_i \cdot B_i \Big/ \left(\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}\right) \quad (4)$$

$$IMD = 1 - norm\left(\sqrt{\sum_{i=1}^{n} (A_i - B_i)^2 \Big/ (\partial_A^2 + \partial_B^2)}\right) \quad (5)$$

A_i and B_i are respectively, components of matrix A and B. At first, the proposed algorithm makes an eleven element vector. The first five vector elements include the sum of similarity amount and inverse distance of the topic sentence with the next following five sentences. Second four elements include the difference between the five previous amounts, respectively. The tenth element include the average of first five primary vector element values and the eleventh element include the average of the second four vector element values respectively (6). In order to reduce the amount of difference gained, determining and applying the difference between the obtained values is important. Therefore, the first sentence in the paragraph is logically removed and the above algorithm is done on the new paragraph with one less sentence. The process repeated (n-5) times (n is the number of paragraph sentences). As result an (n-*11) dimension matrix is created for each paragraph in the document (7).

$$\begin{aligned} V &= \{v_1, v_2, \ldots, v_{11}\} \\ v_{i=1}^{5} &= sum(cst)_i + sum(imdt)_i \\ v_{i=6}^{9} &= v_{i=1}^{5} - v_{i+1}^{6} \\ v_{10} &= median(v_{i=1}^{5}) \\ v_{11} &= median(v_{i=6}^{9}) \end{aligned} \quad (6)$$

To evaluate local coherence at paragraph level, we calculate first six sentences dependency and create the first vector (6). Then, the method calculates other sequential six sentences dependencies to obtain and create paragraph dependency matrix (7). It is concluded that public coherence can be assessed by creating virtual paragraphs. Virtual paragraphs consist of document title and topic sentences of each paragraph. The mentioned method is done on new virtual paragraph to evaluate public coherence.

$$\begin{aligned} M &= \{v_{1,k}, v_{2,k}, \ldots, v_{11,k}\} \\ v_{i=1,k=1}^{i=5,k=n-5} &= sum(cst)_{i,k} + sum(imdt)_{i,k} \\ v_{i=6,k=1}^{i=9,k=n-5} &= v_{i=1,k}^{5} - v_{i+1,k}^{6} \\ v_{i=10,=1}^{k=n-5} &= median(v_{i=1,k}^{5}) \\ v_{i=11,k=1}^{k=n-5} &= median(v_{i=6,k}^{9}) \end{aligned} \quad (7)$$

Below, you can see the proposed algorithm:

1. Calculate the IWMD of first sentence in paragraph (sentence i) and other next five sequential sentences.

a. First sentence and second sentence.

b. .....

c. First sentence and sixth sentence.

d. Generates first Vi vector.

2. Calculate the IWMD of next sentence (sentence i+1) and other next five sequential sentences.

3. Appling n-5 times step 2 (n= number of paragraph sentences)

4. Generate the matrix M.

B. Two Sentences Coherence Evaluation

The two most common ways to show the sentence relentless are by a bag of words (BOW) or based on their term frequency inverse document frequency (TF-IDF). One major drawback of this approach is that the features are completely dependent on words appearance or spelling in the text and often not suitable for evaluating sentences distances or similarity. Capturing distance or similarity of individual words and synonyms is another potential concern, because entity based models makes no attempt to differentiate between different types of synonyms. For more information, look at the following two sequence sentences with different words:

- my master got a hard test
- but my teacher did not give me a good score on this exam

When these sentences have no common words, they convey nearly the same information and cannot be represented by the BOW model. For this case, the closeness of the word pairs: (my, me); (master, teacher); (got, give); and (test, exam) is not factored into the BOW-based distance.

To overcome the mentioned drawbacks, our approach uses recent results by Mikolov et al. [48] whose popular word2vec model generates word embeddings of very large data sets and Inverse Word Mover's Distance (IWMD), utilizes the property of

word2vec embeddings [49]. We represent sentences as a numeric matrix made up of word2vec vectors [51]. The distance between two sequential sentences A and B is the minimum cumulative distance and maximum similarity words from sentence A need to travel to match exactly the words in sentence B.

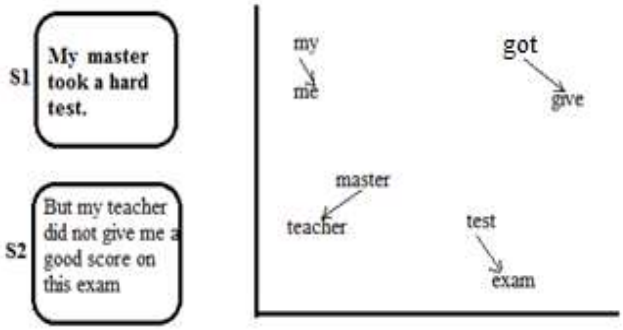

Figure 1: An example of the inverse word mover's distance.

Word Mover's Distance (WMD) is to incorporate the semantic similarity between individual word pairs (e.g. master and teacher) into sentence distance metric. One such measure of word similarity and dissimilarity is naturally provided by their Cosine similarity and inverse Manhattan distance in the word2vec embedding space. The "travel cost" of two words is a natural building block to evaluate distance between two sentences. In this method, first, the similarity between each word in the smaller sentence is compared with each single word in the larger sentence and then the distance of them is calculated. If d1 and d2 be the words representation of two sentences, first, we allow each word $d1_i$ in d1 to be transformed into any word in $d2_j$. Then we obtain word2vec cosine similarity of two sample sentences words (table 1) and normalized word2vec inverse Manhattan distance of them (table 2).

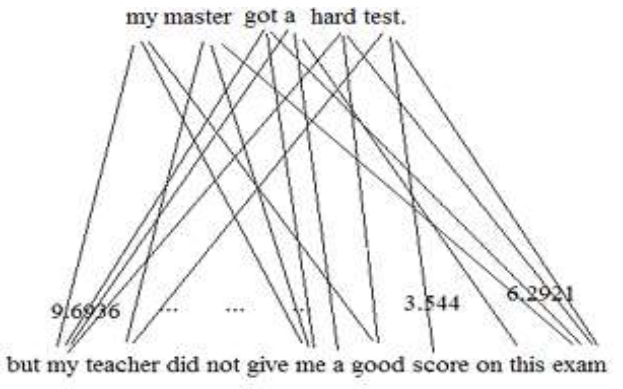

Figure 2: An illustration of cumulative word similarity and inverse word mover's distance.

By cosine similarity table, the three largest word similarities are selected as synonym or nearest embedded words (CST) (table 3) and by inverse normalized Manhattan table, all the same position table 3 values are selected as the best nearest words selected by Manhattan distances (IMDT) (table 4). The IWMD is calculated by summation of two CST and IMDT (table 5). As we show, our method removed statistically most of the stop words based on their importance on analyzed sentences. Sum of matrix values is the two sentence dependency and the first value of vector V_i. The mentioned two sentences are 32.9580.

7. DATA SET

Our database is a data set created by ourselves. We have selected 20 standard Anderson short stories with acceptable coherence from skilled authors. For each document, ten other texts are created by relocating their sentences. Their sentence displacements are 10%, 20% … 100% for the ten generated text.

The other incoherence texts are created by randomly summarized documents. To make incoherence summarized for each document, 30 other texts are created by removing sentences randomly. The thirty texts are made up of ten texts by dropping randomly 10% of the sentences, ten texts by dropping randomly 20% of the sentences and ten texts by dropping randomly 30% of the sentences. These texts are essentially summary texts that do not use any text summarization template. Given the importance of removing sentences, their coherence is decreased. As a result, we have a database of 1020 documents, with different degrees of coherence.

8. EVALUATION

In the current study, comparing an entity based method with word embeddings method shows that the syntactic approaches have much more mean degree than semantic approaches. The yields in this investigation are higher as compared to those of other previous studies and also offer a very simple evaluation.

In order to evaluate the proposed ECEM method, it is compared to Lioma and Tarissan Bipartite Graph Structure of Entity Grids method (BGSEG) [31]. Firstly, we selected one story in the database with other stories included, it is eleven decreasing coherence samples to apply to the model. The results obtained from the preliminary analysis of our method and (BGSEG) method on one of the selected stories are illustrated and compared in Table 6. The selected story has 192 sentences with 4506 words. Test comparison sample includes original text, five examples of the relocated sentence position of 10% to 50%, two examples of 10% randomly summarized text, two examples of 20% randomly summarized text

and one examples of 30% randomly summarized text. As shown in Table 6, there is a difference, about 2.63% improvement, between the results of the two groups, (one is our proposed method and the other is (BGSEG) method), on one selected document.

Table 6 compares the experimental data on one selected document (original story) and its incoherence samples (displacement sentences, randomly summarized). Details of the results obtained on the preliminary analysis of the two methods as shown in Table 6, are set out as follows:

A=selected document

1: Original story

2: 10% displacement sentences

3: 20% displacement sentences

4: 30% displacement sentences

5: 40% displacement sentences

6: 50% displacement sentences

7 and 8: two 10% randomly summarized text. **a** and **b** are two different summaries with different 10% removed sentences..

9 and 10: two 20% randomly summarized text. **a** and **b** are also two different summaries with different 20% removed sentences..

C= the degree of coherence obtained by the (ECEM) compared to the original text

D= the degree of coherence obtained by the (ECEM) compared to B

E= the degree of coherence obtained by the (BGSEG) method compared to the original text

F= the degree of coherence obtained by the (BGSEG) method compared to B

In the next step, the action was performed on ten selected stories and other their decreasing coherence samples with different length and sentences number. From the chart in figure 3 and table 7, it can be seen that our proposed method has a better outcome on long documents with more sentences. Short documents with Low-sentence number have better responded in a graph-based model.

In the current study, comparing our proposed method with BGSEG method showed that the mean degree of coherence evaluation 1.19 percent improvement (table 7). The results in this study also indicate improvement results are much more in larger texts with more sentences.

TABLE 1
OBTAINED WORD2VEC COSINE SIMILARITY OF TWO COMPARING SENTENCES WORDS

	but	my	teacher	didn't	give	me	a	good	score	on	this	exam
my	0.2768	5.1709	0.5004	-0.0049	0.4587	0.8371	0.4467	0.9977	0.0285	3.2963	0.4343	0.7191
master	-0.0261	1.0230	1.1791	-0.0038	0.0722	0.6865	-0.0309	0.3392	1.2155	0.0834	-0.1651	1.8963
got	-0.1050	0.9963	-0.0745	0.0045	0.1534	2.5396	-0.0767	0.4615	0.1767	-0.0023	-0.1902	1.1142
a	0.6095	1.1821	0.0836	0.0110	0.7815	-0.3176	1.9540	0.9347	0.5340	0.8370	1.0101	-0.0962
hard	0.1498	0.9448	0.4168	-0.0022	0.1219	0.1820	0.0892	1.5642	0.6716	0.6330	0.1259	0.7079
test	0.2775	0.0675	0.7443	0.0141	0.4345	0.2035	0.3047	0.3695	1.9082	0.7149	0.4455	1.9731

TABLE 2
OBTAINED WORD2VEC NORMALIZED INVERSE MANHATTAN DISTANCES OF TWO COMPARING SENTENCES WORDS

	but	my	teacher	didn't	give	me	a	good	score	on	this	exam
my	0.1242	1.0000	0.0975	0.0319	0.1358	0.1354	0.1451	0.1905	0.0404	0.0724	0.1470	0.0982
master	0.0336	0.1349	0.1725	0.0331	0.0599	0.1168	0.0365	0.0933	0.1672	0.0575	0.0033	0.1902
got	0.0094	0.1248	0.0414	0.0592	0.0801	0.3264	0.0286	0.1102	0.0545	0.0477	0.0000	0.1171
a	0.2897	0.1451	0.0564	0.0762	0.2279	0.0156	1.0000	0.1819	0.0905	0.1733	0.3591	0.0350
hard	0.0870	0.1296	0.0876	0.0329	0.0664	0.0895	0.0614	0.2925	0.1078	0.1390	0.0719	0.0948
test	0.1440	0.0482	0.1258	0.0787	0.1431	0.0670	0.1237	0.1002	0.2531	0.1569	0.1683	0.1985

TABLE 3
OBTAINED THREE LARGEST WORD SIMILARITIES WORD2VEC COSINE SIMILARITY OF TWO COMPARING SENTENCES WORDS (CST)

	my	teacher	me	a	good	score	this	exam
my	5.1709		0.8371		0.9977			
master		1.1791				1.2155		1.8963
got	0.9963		2.5396					1.1142
a	1.1821			1.9540			1.0101	
hard	0.9448				1.5642			0.7079
test		0.7443				1.9082		1.9731

12: average output.

B=Real percentage of texts coherence

TABLE 4
OBTAINED THE SAME POSITION THREE LARGEST NORMALIZED INVERSE MANHATTAN DISTANCES OF TWO COMPARING SENTENCES WORDS (IMDT)

	my	teacher	me	a	good	score	this	exam
my	1.0000		0.1354		0.1905			
master		0.1725				0.1672		0.1902
got	0.1248		0.3264					0.1171
a	0.1451			1.0000			0.3591	
hard	0.1296				0.2925			0.0948
test		0.1258				0.2531		0.1985

TABLE 5
CUMULATIVE OF THREE LARGEST NORMALIZED INVERSES MANHATTAN AND COSINE SIMILARITY DISTANCES OF TWO COMPARING SENTENCES WORDS

	my	teacher	me	a	good	score	this	exam
my	6.1709		0.9725		1.1882			
master		1.3516				1.3827		2.0865
got	1.1211		2.8660					1.2313
a	1.3272			2.9540			1.3692	
hard	1.0744				1.8567			0.8027
test		0.8701				2.1613		2.1716

TABLE 6
COMPARISON OF THE SUGGESTED METHOD AND (BGSEG) MODEL ON ONE SELECTED STORY AND ITS NON-COHERENT EXAMPLES

	A	B	C	D	E	F
1	Original	100	88	88	86	86
2	10%	90	82	91.11	81	90
3	20%	80	69	86.25	66	82.5
4	30%	70	60	85.71	57	81.43
5	40%	60	58	96.67	57	95
6	50%	50	39	78	37	74
7	10%_a	90	77	85.56	74	82.22
8	10%_b	90	81	90	79	87.78
9	20%_a	80	66	82.5	65	81.25
10	20%_b	80	68	85	66	82.5
11	30%	70	52	74.29	50	71.43
12	Average			85.73		83.1

TABLE 7
COMPARISON OF THE SUGGESTED METHOD AND (BGSEG) METHOD ON TEN STORIES AND ITS NON-COHERENT EXAMPLES

Stories	Number of sentences	BGSEG accuracy	Our method accuracy
ST_45	45	86.22	85
ST_68	68	88.25	88
ST_70	70	86.25	86
ST_82	82	82.11	82.75
ST_86	86	89.5	90.2
ST_101	101	82	83.5
ST_111	111	83.55	85.65
ST_169	169	90.12	92
ST_192	192	83.67	86.85
ST_260	260	86.7	90.35
Average		85.84	87.03

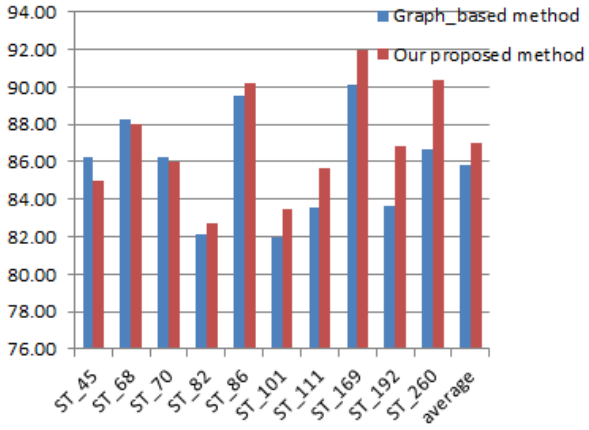

Figure 3: Comparison of the suggested method and (BGSEG) model on ten stories and its non-coherent examples.

9. CONCLUSION

The present study was designed to determine the effect of combining word2vec vectors and most likely n-grams with cohesive LD-n-grams perplexity for representing and measuring text coherence. One of the more significant findings to emerge from this study is statistical framework, evaluating local and global coherence, simultaneously. The second major finding was the proceedings notion of local coherence in paragraph level instead of only few consecutive sentences.

A key strength of our method is the proposed method neither involves words semantic concepts, nor suffers from the computational complexity and data fragmentation. It also has an easier text pre-processing. The result findings of this study rely on

shallow matrix properties and much more inexpensive. It has also gone some way towards enhancing global coherence to all paragraphs relationship in a document instead of individual sentences dependency to title and topic subject. This work can be applied to other languages, if they are provided with word vectors. A further study could assess the long-term effects of our numeric matrix representation of sentences on other NLP tasks without many modifications. Extracted results for sentence reordering documents and also randomly summarized texts show the superiority of the proposed model on long documents. It is suggested that factors such as text summarization, text generation, writer mode detection, topic segmentation, smart CAPTCHA, persian text coherence evaluation, and other text processing fields are further investigated in future studies.

REFERENCES

[1] Z. Lin, C. Liu, H. T. Ng, and M.-Y. Kan, "Combining coherence models and machine translation evaluation metrics for summarization evaluation," in *Proc. The 50th Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 1006–1014, Jeju, Republic of Korea, 2012.

[2] D. Xiong, Y. Ding, M. Zhang, and C. L. Tan, "Lexical chain based cohesion models for document-level statistical machine translation," in *Proc. 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1563–1573, Washington, USA, 2013.

[3] D. Xiong, M. Zhang, and X. Wang, "Topic-based coherence modeling for statistical machine translation," *Trans. Audio, Speech and Lang.,* vol. 23, no. 3, pp. 483–493, 2015.

[4] H. J. Fox, "Phrasal cohesion and statistical machine translation," in *Proc*. *The Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Philadelphia, USA, pp. 304-311, 2002.

[5] H. Yannakoudakis and T. Briscoe, "Modeling coherence in ESOL learner texts," in *Proc*. *The Seventh Workshop on Building Educational Applications Using NLP*, Montreal, Canada, pp. 33-43, 2012.

[6] J. Burstein, J. Tetreault, and S. Andreyev, "Using entity-based features to model coherence in student essays," in *Proc. NAACL-HLT*, California, USA, pp. 681-684, 2010.

[7] D. Higgins, J. Burstin, D. Marcu, and C. Gentile, "Evaluating multiple aspects of coherence in student essays," in *Proc. NAACL-HLT*, pp. 185-192, Boston, USA, 2004.

[8] A. Celikyilmaz and D. Hakkani-Tur, "Discovery of topically coherent sentences for extractive summarization," in *Proc. The 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oregon, USA, pp. 491–499, 2011.

[9] D. Parveen and M. Strube, "Integrating importance, non-redundancy and coherence in graph-based extractive summarization," in *Proc. The Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI*), pp. 1298–1304, Buenos Aires, Argentina, 2015.

[10] R. Zhang, "Sentence ordering driven by local and global coherence for summary generation," in *Proc. The ACL-HLT 2011 Student Session*, Portland, OR, USA, pp. 6–11, 2011.

[11] M. A. K. Halliday and R. Hasan, "Cohesion in English," London, Longman, 1976.

[12] B. J. Grosz, A. K. Joshi, and S. Weinstein, "Centering: A framework for modeling the local coherence of discourse," *Computational Linguistics*, vol. 21, no. 2, pp. 203–225, 1995.

[13] I. Tapiero, "Situation models and levels of coherence: towards a definition of comprehension," Routledge, first edition, ISBN-13: 978-1138004221, 2014.

[14] P. W. Foltz, W. Kintsch, and T. K. Landauer, "The measurement of textual coherence with latent semantic analysis," *Discourse Processes*, vol. 25, no. 2-3, pp. 285-307, 1998.

[15] R. Barzilay and M. Lapata, "Modeling local coherence: An entity-based approach," in *Proc. ACL '05 the 43rd Annual Meeting on Association for Computational Linguistics*, pp.141-148, Michigan, USA, 2005.

[16] R. Barzilay and M. Lapata, "Modeling local coherence: An entity-based approach," *Computational Linguistics*, vol. 34, pp. 1-34, 2008.

[17] S. Somasundaran, J. Burstein, and M. Chodorow, "Lexical chaining for measuring discourse coherence quality in test-taker essays," in *Proc. COLING the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 950–961, Dublin, Ireland, 2014.

[18] V. W. Feng and G. Hirst, "Extending the entity-based coherence model with multiple ranks," in *Proc. EACL*, pp. 315-324, Avignon, France, 2012.

[19] C. Petersen, C. Lioma, J. G. Simonsen, and B. Larsen, "Entropy and graph based modeling of document coherence using discourse entities: An application to IR," in *Proc. ICTIR*, pp. 191–200, Northampton, MA, USA, 2015.

[20] M. Zhang, V. W. Feng, B. Qin, G. Hirst, T. Liu, and J. Huang, "Encoding world knowledge in the evaluation of local coherence," in *Proc. NAACL HLT*, pp. 1087–1096, Denver, Colorado, USA, 2015.

[21] Z. H. Lin, H. T. Ng, and M. Y. Kan, "Automatically evaluating text coherence using discourse relations," in *Proc. ACL-11*, pp. 997-1006, Portland, USA, 2011.

[22] A. Louis and A. Nenkova, "A coherence model based on syntactic patterns," in *Proc. EMNLP-CNLL*, pp. 1157-1168, Jeju Island, Korea, 2012.

[23] R. Iida and T. Tokunaga, "A metric for evaluating discourse coherence based on coreference resolution," in *Proc. COLING*, pp.483-494, Mumbai, India, 2012.

[24] M. Elsner and E. Charniak, "Coreference-inspired coherence modeling," in *Proc. ACL-08*, pp. 41-44, Ohio, USA, 2008.

[25] R. Barzilay and L. Lee, "Catching the drift: probabilistic content models, with applications to generation and summarization," in *Proc. NAACL-HLT*, pp. 113-120, 2004.

[26] M. Elsner, J. Austerweil, and E. Charniak, "A unified local and global model for discourse coherence," in *Proc. NAACL*, pp. 436-443, New York, USA, 2007.

[27] F. Xu, Q. Zhu, G. Zhou, and M. Wang, "Cohesion-driven discourse coherence modeling," *Journal of Chinese Information Processing*, vol. 28, no. 3, pp. 11-21, 2014.

[28] K. Filippova, M. Strube, "Extending the entity-grid coherence model to semantically related entities," in *Proc. ENLG '07 the Eleventh European Workshop on Natural Language Generation*, pp. 139-142, 2007.

[29] M. Lapata and R. Barzilay, "Automatic evaluation of text coherence: models and representations," in *Proc. The 19th International Joint Conference on Artificial Intelligence*, pp. 1085-1090, Scotland, UK, 2005.

[30] F. Xu and S. Du, "An entity-driven recursive neural network model for Chinese discourse coherence modeling," *International Journal of Artificial Intelligence and Applications (IJAIA)*, vol. 8, no. 2, pp. 1-9, 2017.

[31] C. Lioma, F. Tarissan, J. Grue Simonsen, C. Petersen, and B. Larsen, "Exploiting the bipartite structure of entity grids for document coherence and retrieval," presented at the 2nd ACM

International Conference on the Theory of Information, Newark, United States, Sep 2016.

[32] C. Guinaudeau and M. Strube, "Graph-based Local Coherence Modeling," in *Proc. The 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, pp. 93–103, 2013.

[33] M. Mesgar and M. Strube "Graph-based coherence modeling for assessing readability," in *Proc. The Fourth Joint Conference on Lexical and Computational Semantics*, pp. 309–318, Denver, USA, 2015.

[34] R. Soricut and D. Marcu, "Discourse generation using utility-trained coherence models," in *Proc. The COLING/ACL on Main conference poster sessions*, pp. 803-810, Sydney, Australia, 2006.

[35] J. Li and E. Hovy, "A model of coherence based on distributed sentence representation," in *Proc. The 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2039-2048, Doha, Qatar, 2014.

[36] J. Li and D. Jurafsky, "Neural net models for open-domain discourse coherence," in *Proc. The 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 198–209, 2017.

[37] L. Logeswaran, H. Lee and D. Radev, "Sentence ordering using recurrent neural networks," *arXiv preprint arXiv:1611.02654*, Nov 2016.

[38] R. Rosenfeld, "A maximum entropy approach to adaptive statistical language modeling," *Computer Speech & Language*, vol. 10, no. 3, pp. 187–228, 1996.

[39] M. Abdolahi and M. Zahedi, "Text coherence new method using word2vec sentence vectors and most likely n-grams," presented at the 3rd Iranian conference and intelligent systems on signal processing (ICSPIS), Shahrood, Iran, 2017.

[40] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.

[41] R. Johnson and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks," in *Proc. The 2015 Annual Conference of the North American Chapter of the ACL*, pp. 103–112, Denver, USA, 2015.

[42] R. Johnson and T. Zhang, "Semi-supervised convolutional neural networks for text categorization via region embedding," in *Proc. Advances in Neural Information Processing Systems (NIPS 2015)*, pp. 919-927, 2015.

[43] T. H. Nguyen, R. Grishman, "Relation extraction: Perspective from convolutional neural networks," in *Proc. Workshop on Vector Space Modeling for NLP at NAACL 2015*, pp. 39–48, Denver, USA, 2015.

[44] N. Kalchbrenner, E. Grefenstette, and P. Blunsom "A convolutional neural network for modelling sentences," in *Proc. The 52nd Annual Meeting of the Association for Computational Linguistics Acl*, pp. 655–665, Baltimore, USA, 2014.

[45] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. The 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, Doha, Qatar, 2014.

[46] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," *in Proc. The 8th International Joint Conference on Natural Language Processing*, pp. 253-263, Taipei, Taiwan, 2017.

[47] F. Yaghmaee and M. Kamyar, "Introducing new trends for persian CAPTCHA," *Journal of Electrical and Computer Engineering Innovations (JECEI)*, vol. 4, no. 2, pp. 119-126, 2016.

[48] T. Mikolov and I. Sutskever, "Distributed representations of words and phrases and their compositionality," in *Proc. NIPS 2013*, pp. 3111–3119, Nevada, USA, 2013.

[49] J. Pennington, R. Socher, and C. D. Manning, "GloVe: global vectors forword representation," in *Proc. The 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, 2014.

[50] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger, "From word embeddings to document distances," in *Proc. The 32nd International Conference on Machine Learning, JMLR: W&CP* vol. 37, pp. 957-966, Lille, France, 2015.

[51] M. Abdolahi and M. Zahedi, "Sentence matrix normalization using most likely n-grams vector," presented at the 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), Tehran, Iran, 2017.

BIOGRAPHIES

Mohamad Abdolahi was born in Mashhad, Iran, on October 22, 1964. He is an Ph.D. candidate in Shahrood University of Technology in the field of computer engineering - artificial intelligence. His is lecturer in Iranian Academic Center for Education, Culture and Research (ACECR), Mashhad, Iran. His special fields of interest are NLP, data mining, image processing and machine learning.

Morteza Zahedi graduated from the RWTH-Aachen University, Aachen, Germany and he is an assistant Professor in Shahrood University of Technology. His special fields of interest are NLP, pattern recognition, image and video processing.

How to cite this paper:
M. Abdolahi and M. Zahedi, "A new model for text coherence evaluation using statistical characteristics," *Journal of Electrical and Computer Engineering Innovations*, vol. 6, no. 1, pp. 15-24, 2018.
DOI: 10.22061/JECEI.2018.799
URL: http://jecei.sru.ac.ir/article_799.html