Journal of Advances in Computer Research
Quarterly ISSN: 2008-6148
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 11, No. 4, November 2020), Pages: 23-37
www.jacr.iausari.ac.ir

An Optimal Approach to Local and Global Text Coherence Evaluation Combining Entity-based, Graph-based and Entropy-based Approaches

Mohamad Abdolahi *
University of Applied Science and Technology, Iranian Academic Center for Education, Culture and Research (ACECR), Khorasan Razavi Branch, Mashhad, Iran
mabdolahi512@yahoo.com

Received: 2020/05/27; Accepted: 2020/10/03

Abstract

Text coherence evaluation becomes a vital and lovely task 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 with nouns and noun phrases change role in sequential sentences within short part of a text. They even have limitations in global coherence evaluation, especially in long and narrative documents. This paper presents a new and simplified method for evaluating local and global text coherence. The proposed method focuses on entity grid method and employs two graph-based and entropy-based approaches to overcome its challenges and shortcomings. Applying statistical approaches, the presented method studies how to incorporate other entity properties into short and long stories to assess both local and global coherence, simultaneously. Results indicate that the proposed method is superior to other algorithms in terms of performance, accuracy in long documents with a high number of sentences.

Keywords: Local coherence, Global coherence, Graph-based coherence, Entity-based coherence, Entropy-based coherence.

1. Introduction

With the enormous amount of machine generated text every day, it is difficult to evaluate their coherence and topic dependency through the manual means. The World Wide Web provides a vast amount of content in forms of web pages, news articles, emails, and access to the databases around the world. However, much of these contents may not be in ordinal manner and difficult to read and understand and also fake websites [1]. The vast number of documents and most valuable information in each text lead to consuming them as a big data. Big data referred to huge datasets with high number of objects and high number of dimensions [2]. Mining and extracting valuable text features is one of the most interesting fields in all approaches of text processing.

Accurate topic understanding has been an interest issue for all researchers in human linguistic communication like text processing, voice processing and sign language. That is, in spoken languages, the coherence of successive sections of text and voice, and in sign languages balanced and consecutive hand movements. Appropriate machine-learning methods are nowadays suggested for understanding the concepts presented for sign languages [3]. Given that text production or text information retrieval is a humans or smart machines feature, therefore, the intelligence of a machine can be evaluated from its coherent textual production or through coherent and relevant text CAPTCHA [4].

Discourse Coherence Modeling (DCM) or Discourse Coherence Evaluation (DCE) tries to evaluate or if possible, to improve the degree of coherence among sequential sentences and paragraph within a whole text. Discourse coherence is an important aspect of text quality and a key property of any well-organized text. Coherence of text is the logical organization and development of thematic content in a document. It evaluates the degree of logical and global consistency for text and tried to convert random set of sentences to consistent order of logically coherent document. It is considered one of the key problems in all Natural Language Processing (NLP) along of its wide usage in many NLP applications, such as statistical machine translation [5], [6], [7], [8], [9], natural language generation (NLG) [10], [11], [12] , automatic text summarization [13], [14], [15], narrative text summarization and information extraction [16], automated essay evaluation [17], [18], [19], [20], question answering [21], [22], text similarity measures [23], sentiment analysis [24], social network structures [25], text classification [26], [27], [28], [29] and text categorization [30]. Thus, all of the text processing systems attempt to assess their output texts and if necessary, improve them. There are some differences between document coherence and document cohesion. Cohesion refers to internal properties of a text, whereas coherence refers to its contextual properties. In the other hand, coherence is the way in which all part of document is related to each other and make a sense in the situation in which occurs. The problem of incoherent text has been addressed from many different perspectives. Through this paper, we have discussed some of novel approaches for automatic evaluation of text coherence. According to Halliday and Hasan [31], in general, to text coherence evaluation generally has many similar components such as lexical overlap or co-reference among all sentences within a text, while incoherent discourse is the sequential random ordering sentences one. The first and important cohesion theories were centering driven [32] and entity-based model [33], [34] which tried to capture the syntactic and semantic distribution of discourse entities (nouns) between two adjacent sentences in a text. Afterwards, many developed works were presented such as multiple ranking model [35], discourse relation-based approach [36] and syntactic patterns-based model [33]. However, the traditional drawback and potential issue for developing them in coherence evaluation and development models need feature engineering.

Document coherence evaluation is divided into two main classes of local and global coherence evaluation. Local coherence is the well conceptual connectedness of adjacent and consecutive sentences through lexical cohesion [31] or entity repetition [37]. The mentioned class attempt to capture the similar content of adjacent sentences and therefore tend to contain related words. Local models are often good at finding sentences that belong near one another in the document. However, they have trouble to finding the similar content at the beginning or end of the document. Another problem is recovering related coherence features from sudden shifts in topic such as occur at paragraph boundaries. Some local models also have trouble deciding which of a pair of related sentences ought to come first. But global coherence is the topic-level relation connecting remote sentences or adjacent and consecutive paragraphs [37]. The main challenge faced by many previous experiments is evaluating the coherence simultaneously on both levels of local and global. Another criticism of much of the proposed methods is their weaknesses in long documents [38].

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 entity grid, graph based and entropy-based methods were used in this investigation.

In this paper, to evaluate text coherence, we propose a paragraph coherence level task as well as conducting sentence ordering. Paragraph level coherence evaluation is one of the most important and favorite approaches in most recent researches [39]. This study provides a combined three previous approaches of Entity-based, Graph-based and Entity&Graph-based on different length stories. Due to upper triangular matrix, our proposed method is considered and measured the relationship between any sentences contained in the paragraph with all other sentences in the document. As result the proposed method evaluate document coherence with no dependency on subject and special field. In light of recent events in statistical methods, the proposed method has some advantages. The mentioned method in this paper attempts to shows suggested approach does not engage with words semantic meanings and no dependency on specific field’s knowledge. The model converts each paragraph into much more compact numerical upper triangular matrix with simple computational algorithm than all other previous methods like entity grid and graph-based methods. These advantages make the proposed model suitable for high redundancy documents, very close sentences concept texts and long narrative documents.

At first, preprocessing algorithm is taking place on our combined proposed method. Then the text is partitioned into its main components (paragraphs, sentences and words). In the next step, using the entity grid algorithm, each paragraph is converted into an upper triangular matrix. In addition to more speed and simple algorithm, the most interesting finding of the method is producing higher accuracy coherence evaluation on more long narrative documents and stories.

The remaining part of the paper proceeds as follows: At first gives a brief overview of the recent history of text coherence evaluation. Section three begins by laying out the text preprocessing methods. The fourth section is concerned with the methodology used for this study. The fifth section presents the findings of the research, and finally conclusion is drawn in section six.

2. Related works

Generating of coherent text has been considered from the beginning introduction of the text summarization technique by Luon and Hans Peter (1958) [40]. But increasing text size, more powerful search engines, achieving and combining facilities of large collection of text resources on web and much more rapid progress of all NLP tasks, have converted coherence and topic dependency of generated texts to one of the most important researchers concerns. The first computerized method of text coherence evaluation was started by Fultz et al (1998) [41]. In their view, a text is coherent when its two successive sentences have semantic connection. They also introduced a lexical meaning vector-based model to calculate successive sentences semantic relation. Since then, a lot of approaches were presented based on their studies. Most of the next proposed approaches are supervised methods and used the amount of thematic relation between successive sentences. In the following the research provides a brief overview of the most important approaches, their progression, classifying other researcher finding and identifying ambiguities in document coherence evaluation.

Entity-based tried to consider local coherence and have a long tradition within the linguistic and cognitive science literature. It is also one of the well-established forms that analyses adjacent sentence's words, to extract their coherence patterns by their grammatical roles [42]. Barzilay and Lapata proposed an entity-based coherence model to evaluate local coherence documents [33], [43], but some other combined novel approaches such as neural network models [44], [45] and original bipartite graph [46] are proposed in recent years. Their model is one of the most important and much applied approaches in document coherence evaluation. In this model, in addition to other previous features, the grammatical role of nouns and nouns phrases in successive sentences is considered as prominent feature. The model learns distribution patterns of nouns grammatical role transition between two adjacent sentences. The most their interesting finding was that a coherent document has regular patterns in changing entities in adjacent sentences. They assumed subject entities as (S), object (O), other (X), and missing (-). They also generate a two-dimensional matrix with rows as sentences and columns showing the entities. In entity-based hypothesis, the distribution of cohesive factors in matrix defines the rules of adjacent sentences coherence. For example, cohesive texts have high-density columns or columns with more amounts of subject (S) and object (O) entities. To measure local coherence, the model computes the degree of semantic relatedness between sentences by taking the mean of all individual transitions (1).

$$\text{Coherence}(d) = \frac{\sum_{i=1}^{n-1} \text{Sim}(S_i, S_{i+1})}{n-1} \tag{1}$$

Where sim (S_i, S_{i+1}) is a measure of similarity between sentences S_i and S_{i+1}. The model also has experimented with three broad classes which uses word-based, distributional, and taxonomy-based similarity measures. For example, word overlap measurement semantic similarity can be operationalized in form of Equation (2):

$$\text{sim}(S_1, S_2) = \frac{2\left|\text{words}(S_1) \cap \text{words}(S_2)\right|}{\left(\left|\text{words}(S_1)\right| + \left|\text{words}(S_2)\right|\right)} \tag{2}$$

In this equation words (S_i) is the set of words in sentence i. The main shortcoming of this measure is low coherence for sentence pairs with no words in common, even though they may be semantically related. Essay scoring is other scope that uses entity-based method. The study by J. Burstein offers combined entity-based features with aspect related to grammar errors to improve the efficiency of automated student essays coherence [18]. The big shortcoming of primary entity-based method was the repeating of real form of nouns in successive sentences. M. Zhang et al. (2015) studied the effects of other words relation such as synonym, antonym, hypernym and hyponym to detect coherence features [47]. They showed in 42.23%, two adjacent sentences did not have common noun entity. Hence, they have suggested mentioned features with semantic connections such as (car, automobile), or (car, petrol station communications). To implement their own theory, they transform the entity grid into sentence graph and measures text coherence by computing the average out-degree of the graph. They also constructed bipartite graph G = (V_s, V_e, L, W), where V_s is the set of nodes representing sentences in the text, V_e is the set of nodes representing entities, L is the set of edges associated with a weight w ϵ W. Then, the bipartite graph G is converted to another graph P that consists of only sentence nodes and its edge connected two sentence nodes if and only if at least one entity is shared between these two sentences. The weight of each edge (P) is computed by aggregating the edge weights in the original bipartite graph G (3):

$$w^{(p)}\left(S_x, S_y\right) = \sum_{e \in E_{xy}} w^{(G)}\left(e, S_x\right) \times w^{(G)}\left(e, S_y\right) \tag{3}$$

In Equation (3) Exy is the shared entities of two sentences S_x and S_y, and w(G)(e, S_x) is also the weight of edge between entity e and sentence S_x.

The entity-based method becomes one of the most popular methods for coherence evaluating, some studies have been done and many scholars have used and tried to improve it. However, it has some weakness that lead to researchers' attention to other areas such as graph theory, neural networks, deep learning, and other combine methods. The most important of their limitation are as follows:

- These methods are suitable for local coherence evaluation. Entity-based methods which evaluate global coherence do not have acceptable accuracy.
- They are able to consider adjacent sentences coherence. Long distance sentences may also have evidence of coherence verification. But the method cannot extract them.
- In these approaches, it is not possible to develop a methodology to provide a model that can offer a way to improve coherence.
- Entity-based models have strong dependency on language and subject matter. Applying them to different languages and subjects, the model's accuracy is greatly reduced.
- They are using semantic entities (Subject and object). Dependency on semantic entities and grammatical rules, the method engages with dimensional complexity. For example, if K is the transition state and R is the grammatical role, the generated network has KR transition [45]. That's why all previous models have used K <= 3.
- The methods for selecting and extracting these entities (subject and object) are completely semantic.
- Compared to statistical methods, they have more computational complexity
- Extract features (subject, object, etc.) make limited feature search scope. The proposed model does not provide a way to extract and use other entities (verbs, attributes, adverbs, etc.) [48].
- The method has high level of data dispersion. Existing entities are sparse throughout all the text, which may be far apart. The proposed method does not provide a solution for finding the dependency of far sentences.
- This approach is useful for short or medium documents. But in dealing with long documents, the speed and precision are diminished.

There are several possible explanations and usage for graph theory on many NLP tasks. However, most of previously mentioned methods of entity-based and graph theory suffer from some serious disadvantages. These results would seem to suggest that some researches focus on combining graph and entity-based method. In order to overcome entity-based problems C. Guinaudeau and M. Strube tried to represent entity-grid into a graph format using a bipartite graph. They claim their combined method avoids data sparsely and some other drawback in graph structure [49]. The graph tracks the presence of all entities as graph nodes and their connections as graph edges. The proposed model measured text coherence by calculating the average of out degree projection and summing the shared edges. Established graph can easily cover the whole text without problems such as computational complexity and data sparsely. Equation (4) is shown the general form of coherence score assigned to a document in this approach [48].

$$S(D) = \frac{1}{N}\sum_{i=1}^{N} O(S_i) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=i+1}^{N} w_{i,j} \tag{4}$$

Where N is the number of sentences represented in the document graph. $O(S_i)$ is the total weight of edges leaving that sentence. This weight is the sum of the contributions of all edges connecting Si to any $S_j \in D$.

Petersen and Simonsen presented model is a combination of graph theory and entropy method for assessing the consistency of document sentences [50]. The novelty of their model is that 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 based on frequent subgraphs were introduced by M. Mesgar [51]. The main idea of the model is that the coherence texts are consistent of particular patterns in their extracted subgraphs.

Graph-based methods have solved the problem of limiting in sequential sentences dependency and local coherence. But they did not propose any acceptable method to declare the type of existing entity and only specify the existence of common entities between the two sentences. The matrix values created by the entity grid method, converted to graphs mapped in graph-based methods. The mapped graph only specifies the existence and the number of common entities between the two sentences. It also does not have the ability to specify the grammatical role of an entity and its frequency in each sentence. In entropy-based approaches, entity types are not a metrics for coherence measurement. They only considered entity types frequencies in adjacent sentences.

Recently, neural network approaches have achieved state-of-the-art results in text coherence evaluation modeling. However, the most of neural network models often fail on harder tasks that they sensitive in local contexts such as candidate ranking in conversational dialogue and machine translation. H. C. Moon and colleague proposed a new coherence model with sentence grammar properties, inter-sentence coherence relations, and global coherence patterns into a common neural framework [52].

One of most famous and common vector and matrix-based text processing is Google word2vec algorithm [53]. The popular method generates an n elements vectors (-1 to 1) that they distributed numerical representations of the context of individual word features (5).

$$\frac{1}{T}\sum_{i=1}^{T}\sum_{j\in nb(t)} \log p(w_i \mid w_t) \tag{5}$$

Where nb(t) is the set of neighboring words and the $p(w_i \mid w_t)$ is the maximum likelihood of travel cost of word w_i to word w_t . It makes highly accurate guesses about a word’s meaning based on its past appearances. We use the word2vec algorithm to evaluate text coherence [37] and coherent text summarization [54] in other previous researches.

Both our combined method and approaches that use word2vec algorithm tried to convert sentences to a numerical matrix. While word2vec method uses neural networks and billions of words within several million web documents to generates a fixed size vector for each word in document, but our proposed method uses a simple upper triangular matrix to evaluate all sentences in paragraph with low cost in time and memory. In compare to word2vec methods our simple method doesn't need to apply high level text preprocessing and more suitable for stories narrative document.

3. Proposed method

Until now, combinatorial techniques have been used in many research areas, especially image processing [55]. But the use of combinatorial techniques is less commonly in word processing, which in this study we refer to it. Experience in using combinatorial techniques in image processing has been demonstrated that in these methods overlapping the advantages of each method reduces the shortcomings of the other methods.

A big challenge with traditional entity grid approaches was their limitation to evaluate two sequential sentences topic dependency and local coherence evaluation. These approaches do not have the ability to assess the coherence of more distance sentences. Let D = (s_1, …, s_n) be a paragraph consisting of n sentences. Our proposed approaches overcome the problem by general comparison of n sentences throughout the D paragraph (6).

$$D = \{s_1, s_2, s_3, \ldots, s_n\} \tag{6}$$

Our proposed method is considered and measured the relationship between any sentences contained in the paragraph with other sentences (7).

$$d = \begin{bmatrix} \{(s_1, s_2), (s_1, s_3), \ldots, (s_1, s_n)\}, \\ \{(s_2, s_3), (s_2, s_4), \ldots, (s_2, s_n)\}, \\ \ldots, \\ \{(s_{n-1}, s_n)\} \end{bmatrix} \tag{7}$$

The created vectors form an upper triangular matrix, each containing the number of common entities between two sentences. Consider following example illustrates the method. Figure (1) shows an example paragraph in the text that used by Simonsen & Petersen approach [47]:

The old **man** said: his **hope** and his **confidence** had never gone.
The **boy** said.
The old **man** agreed: **you** didn't steal **them**?
The **boy** said: but **I** bought **these**.
Thank **you**, the old **man** said.

Figure 1. An example paragraph of fictional text

Table (1) illustrated the upper triangular matrix five sentences mentioned paragraph with common entities.

Table 1. Upper triangular matrix created from common entities in five sentences paragraph

D	S1	S2	S3	S4	S5
S1		0	1	0	1
S2			0	1	0
S3				0	1
S4					0

In first view, the higher of mean values of the upper triangular matrix, considered as high global coherence of paragraph. The generated primary matrix is an elementary matrix that does not consider many coherent factors between the two sentences. It is only attending the number of common entities among sentences. In the present study, other coherent factors are attended to determine the numerical value as well as improving the upper triangular matrix. The other coherent factors considered in the proposed approach are the following:

- In our approach, the types of entities (subject, object, etc.) are affected on scoring integrity of two sentences.
- In original entity grid approach, if an entity had two grammatical roles, always the role with higher priority was attended. But in our proposed approach, both roles will be considered.
- In original entity grid approach, if an entity was repeated more than once in a single role, one score was given to it. But in our proposed approach, more scores are given to coherence of two sentences with more role common entities.
- In previous methods, far space sentences with common entity have no effect on coherence evaluation, but in our proposed method, far sentenced common entities are attended.

Lower variation of entities in terms of word and grammatical position is a sign of higher coherence of the text. Based on Simonsen & Petersen theory, more entity variations in a text specifies the focus of text on more subjects and thus less text coherence [50]. Hence, increasing the average number of entities in the matrix and decreasing the mean of entity's diversity, the text has higher local and global coherence. To interference the entity type, entity number, two sentences distance and its frequency in paragraph coherence evaluation, we propose some simple and novel solutions.

In our model, a constant score is considered for every entity based on its grammatical role in sentence. Such as graph-based approaches, a fixed number is defined for all entities. As result “3” for subject entity, “2” for object entity and “1” for others are considered. The upper triangular matrix elements are calculated by Equation (8):

$$\mathrm{Score}\left(S_i, S_{i+1..n}\right)_{i=1..n} = G\_Pos\left(S_i\right) \times G\_Pos\left(S_{i+1..n}\right) \times \frac{1}{\mathrm{Dist}\left(S_i, S_{i..n}\right)} \quad (8)$$

Where Score (S_i, $S_{i+1..n}$) is two sentences coherence score, G_Pos(S_i) is the elected number of entity grammatical role in first sentence, G_Pos($S_{i+1..n}$) is the elected number of entity grammatical role in second sentence and Dist(S_i, $S_{i..n}$) is the two sentences distance.

Entity grid approach only considers two adjacent sentences relationships. In this approach, non-neighboring sentences have no relation and their relationship score is zero. In the graph-based approaches, the relationship between far sentences is considered. But in long documents the generated graph is very complex.

We used a kind of sentence mover's similarity method [56] that find all relation properties in all sentences paragraph. Our proposed evaluated metrics is inverse distance criteria as new metric that evaluate two far sentence coherence. For example, in two non-adjacent sentences and first sentence with subject entity (3), second sentence with object entity (2) and the distance between two sentences is 3; their coherence score is calculated as follow:

$$2= 3* 2* 1/3$$

Then paragraph's upper triangular matrix is generated according to calculated two sentences coherence score. By production of inverse distance criteria value in other values, we have small numerical coherence score for two far sentences. In this case the averaging matrix elements becomes a small amount and it will have a reverse effect on coherence degree of paragraph. In this case we use the normalized sum of matrix values to score paragraph coherence. Equation (9) shows the normalized our paragraph coherence score:

$$\text{Coherence}(d) = \frac{\sum S(i,j)_{i=1..n,\, j=i..n}}{\text{Opt\_Coh}(d)} \tag{9}$$

Where Coherence(d) is the paragraph coherence score, $\sum S(i,j)_{i=1..n,\ j=i..n}$ is the sum of matrix values and Opt_Coh(d) is best optimal matrix values. The optimal matrix value is the value obtained when all entities in two sequential sentences are in the same grammar position.

4. Evaluation of proposed method

The following example illustrates the method. We apply our method on simple figure (1) example paragraph. So, at first step, as in the entity grid approach, entities are selected, extracted and their entity matrix is formed.

Table 2. Entity matrix of selected text sentences

man	hope	Confidence	boy	you	them	I	these
S	S	S	_	_	_	_	_
_	_	_	S	_	_	_	_
S	_	_	_	S	O	_	_
_	_	_	S	_	_	S	O
S	_	_	_	O	_	_	_

In this matrix, (S) value denotes as subject entity, (O) as object entity and (-) is other entity. In the second step, using the entity matrix and equation (8), the upper triangular matrix is formed. For example, the first row of paragraph upper triangular matrix {(S_1, S_2), (S_1, S_3), (S_1, S_4), (S_1, S_5)} will be set:

man (1, 2) = 3*1*1=3
hope (1,2) =3*1*1=3

confidence (1,2) =3*1*1=3
Sum (man (1,2), hope (1,2), confidence (1,2)) = 9

man (1,3) = 3*3*1/2=4.5
hope (1,3) =3*1*1/1=1.5
confidence (1,3) =3*1*1/1=1.5
Sum (man (1,3), hope (1,3), confidence (1,3)) = 7.5

man (1,4) = 3*1*1/3=1
hope (1,4) =3*1*1/3=1
confidence (1,4) =3*1*1/3=1
Sum (man (1,4), hope (1,4), confidence (1,4)) = 3

man (1,5) = 3*3*1/4=2.25
hope (1,5) =3*1*1/4=0.75
confidence (1,5) =3*1*1/4=0.75
Sum (man (1,5), hope (1,5), confidence (1,5)) = 3.75

In this way, other values are calculated and finally the upper triangular matrix is created.

Table 3. Selected paragraph upper triangular matrix coherence

D	S1	S2	S3	S4	S5
S1		9	7.5	3	3.75
S2			3	1.5	1
S3				9	9
S4					9

To evaluate the paragraph coherence, we calculate the ratio of highest possible state (the state where all entities are the subject) on the sum of the matrix values.

(9+7.5+3+3.75+3+1.5+1+9+9+5) / 10=5.175

5. Experimental results

To evaluate the impact of local and global coherence, we conduct experiments on 20 standard selected Anderson stories with acceptable coherence by skilled authors. To generate incoherent text, we create a set of texts with different degrees of coherence by relocating and removing randomly sentences. For each story, ten texts are created by relocating their sentences 10%, 20% … 100% and to generate randomly summarized text, we remove randomly 10%, 20%, 30%, 40% and 50% sentences to create five other incoherent texts. As a result, we have a database of 320 documents, with different degrees of coherence. In our experiments, we use the Stanford parser to automatically extract the grammatical role for each entity mention. Table (4) shows the performance of various models on short and long stories.

Table 4. Accuracies of various four models on the short and long stories

Models	Short stories (40-100 sentences)	Long stories (100-300 sentences)
Entity-based	85.9%	80.2%
Graph-based	87%	81%
Entity & Graph-based	88.9%	82.2%
Entity & Graph & Entropy-based	90.02%	91.3%

Table (4) provides the results obtained from the preliminary analysis of Entity-based, Graph-based and Entity & Graph-based on short stories (40-100 sentences) and long stories (100-300 sentences). What is interesting about the results in this table is that all three models have acceptable result on short documents, while by increasing the document length, our combined proposed method tends to has better result on long stories. The graph in figure 2 shows that there has been a slight increase in the obtained accuracy of our method on the documents with more sentences than other previous models.

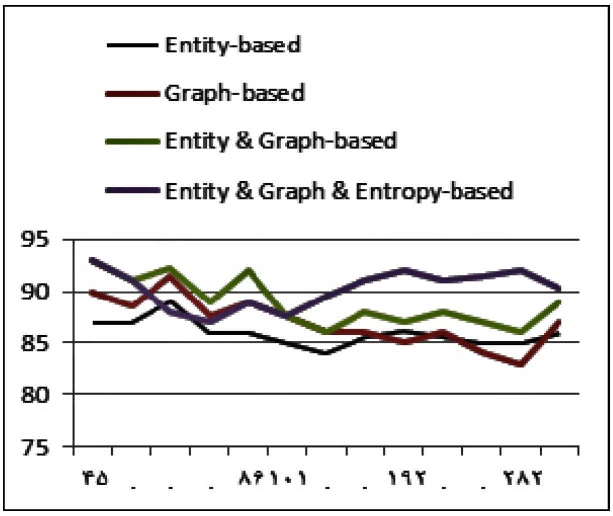

Figure 2. Accuracies of various four models on the short and long stories

6. Conclusion

The present study was designed to determine the effect of combining entity grid model, graph-based model and entropy-based model to represent and measure text coherence. The second aim of this study was to investigate the effects of statistical frameworks and evaluating local and global coherence simultaneously. Another important finding was the acceptance of local coherence in paragraph level instead of only few consecutive sentences. The combination of our method findings provides the beneficial of three novel and famous text coherence methods. An implication of this study is much easier of its text pre-processing algorithm. The result findings of this study rely on shallow triangular top-matrix properties and much more inexpensive. One unanticipated finding of our method was enhancing global coherence to all document's paragraphs relationship instead of individual sentences dependency to title and topic subject. The principal theoretical implication of this study is that it can be applied to other languages, if they are provided with noun and noun phrases WordNet. A further study could assess the long-term effects of our combined triangular top-matrix representation of paragraphs on other NLP tasks without many modifications. Experiments show that our proposed model has better performance than the state-of-the-art systems on long and narrative documents. Further research needs to examine more closely the links between text summarization and its local and global coherence evaluation.

References

[1] F. Parandeh, F. Motlagh, and A. Bardsiri, "Detecting Fake Websites Using Swarm Intelligence Mechanism in Human Learning," International Journal of Engineering (IJE), IJE TRANSACTIONSI, vol. 31, no. 10, pp. 1642-1650, 2018.

[2] I. Behravan, S. H. Zahiri, S. M. Razavi, and R. Trasarti, “Clustering a big mobility dataset using an automatic swarm intelligence-based clustering method,” Journal of Electrical and Computer Engineering Innovations, vol. 6, no. 2, pp. 243-262, 2018.

[3] A. Safaei, and M. Jahed, "3D Hand Motion Evaluation Using HMM", Journal of Modern Education and Computer Science (MECS), vol. 1, no. 1, pp. 11-18, 2013.

[4] F. Yaghmaee, and M. Kamyar, "Introducing New Trends for Persian CAPTCHA", Journal of Modern Education and Computer Science (MECS), vol. 4, no. 2, pp. 119-126, 2016.

[5] H J. Fox, "Phrasal Cohesion and Statistical Machine Translation," in Proc. of the ACL-02 conference on Empirical methods in natural language processing (EMNLP, Vol. 10, pp. 304-311, 2002.

[6] Z. Lin, C. Liu, H. T. Ng, and M. Y. Kan, “Combining coherence models and machine translation evaluation metrics for summarization evaluation," in Proc. of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers., vol. 1, pp. 1006-1014, 2012.

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

[8] O. Bakay, B. Avar, and O.T. Yildiz, A tree-based approach for English-to-Turkish translation. Turkish Journal of Electrical Engineering & Computer Sciences, vol. 27, no. 1, pp. 437-452, 2019.

[9] H. Xiong, H. He, H. Wu, H. Wang, "Modeling coherence for discourse neural machine translation", Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 7338-7345, 2019.

[10] S. Soricut, and D. Marcu, “Discourse Generation Using Utility-Trained Coherence Models," in Proc. of COLING-ACL, pp. 803-810, 2006.

[11] R. Barzilay, and L. Lee, " Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization," in Proc. of NAACL-HLT, pp. 113-120, 2004.

[12] A. Celikyilmaz, E. Clark, J. Gao, "Evaluation of Text Generation: A Survey", arXiv preprint arXiv:2006.14799, 2020.

[13] A. Celikyilmaz, and D. Hakkani-Tur, "Discovery of topically coherent sentences for extractive summarization," in Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies., vol. 1, pp. 491-499, 2011.

[14] D. Parveen, and M. Strube, "Integrating Importance, Non-Redundancy and Coherence in Graph-Based Extractive Summarization," in IJCAI, vol. 31, no. 10, pp. 1298-1304, 2015.

[15] R. Zhang, "Sentence ordering driven by local and global coherence for summary generation," in Proc. of the ACL 2011 Student Session: Association for Computational Linguistics, pp. 6-11, 2011.

[16] L. Ermakova, J. V. Cossu, J. Mothe, "A survey on evaluation of summarization methods", Information Processing & Management, vol. 56, no. 5, pp. 1794-1814, 2019.

[17] H. Yannakoudakis, and T. Briscoe, "Modeling coherence in ESOL learner texts," in Proc. of the Seventh Workshop on Building Educational Applications Using NLP: Association for Computational Linguistics, pp. 33-43, 2012.

[18] J. Burstein, Tetreault., and S. Andreyev, " Using entity-based features to model coherence in student essays," in Human language technologies: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics, pp. 681-684, 2010.

[19] D. Higgins, J. Burstein, D. Marcu, and C. Gentile, "Evaluating multiple aspects of coherence in student essays," in Proc. of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL, 2004.

[20] H. K. Janda, A. Pawar, S. Du, V. Mago, "Syntactic, Semantic and Sentiment Analysis: The Joint Effect on Automated Essay Evaluation", IEEE Access, vol. 7, pp. 108486-108503, 2019.

[21] M. Jamehshourani, A. Fatemi, and M. Nematbakhsh, "Implicit relation based question answering to answer simple questions over DBpedia ", Turkish Journal of electrical engineering & computer sciences, vol. 28, no. 3, pp. 1474-1490, 2020.

[22] H. Sadr, M. Nazari Solimandarabi, "Presentation of an Efficient Automatic Short Answer Grading Model Based on Combination of Pseudo Relevance Feedback and Semantic Relatedness Measures", vol. 10, no. 2, pp. 17-30, 2019.

[23] H. Mohammadi, and S.H. Khasteh, A Fast Text Similarity Measure for Large Document Collections using Multi-reference Cosine and Genetic Algorithm. arXiv preprint arXiv:1810.03102, 2018.

[24] M. T. Ikram, N.A. Butt, and M.T. Afzal, Open source software adoption evaluation through feature level sentiment analysis using Twitter data. Turkish Journal of Electrical Engineering & Computer Sciences, vol. 24, no. 5, pp. 4481-4496, 2016.

[25] M. S. Cinar, B. Genc, and H. Sever, "Identifying criminal organizations from their social network structures", Turkish Journal of Electrical Engineering & Computer Sciences, vol. 27, no. 1, pp. 421-436, 2019.

[26] A. H. Jadidinejad, V. Marza, "Building Semantic Kernel for Persian Text Classification with a Small Amount of Training Data", Journal of Advances in Computer Research, vol. 6, no. 1, pp. 125-136, 2015.

[27] M. Rostami, S. Ayat, I. Attarzadeh, F. Saghari, "Processing a method to classify texts using data mining", Journal of Advances in Computer Research, vol. 6, no. 4, pp. 125-137, 2015.

[28] H. Majidpour, F. Soleimanian Gharehchopogh, "An Improved Flower Pollination Algorithm with Ada Boost Algorithm for Feature Selection in Text Documents Classification", Journal of Advances in Computer Research, vol. 9, no. 1, pp. 29-40, 2018.

[29] A. Allahverdipoor, F. Soleimanian Gharehchopogh, A"n Improved K- Nearest Neighbor with Crow Search Algorithm for Feature Selection in Text Documents Classification", Journal of Advances in Computer Research, vol. 9, no. 2, pp. 37-48, 2018.

[30] A. Allahverdipoor, F. Soleimanian Gharehchopogh, "A New Hybrid Model of K- Means and Naïve Bayes Algorithms for Feature Selection in Text Documents Categorization", Journal of Advances in Computer Research, vol. 8, no. 4, pp. 37-48, 2018

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

[32] B. J. Grosz, S. Weinstein, and A. K. Joshi, "Centering: A Framework for Modeling the Local Coherence of Discourse," Computational Linguistics, vol. 21, no. 2, pp. 203-225, 1995.

[33] R. Barzilay, and M. Lapata, " Modeling local coherence: an entity-based approach," in Proc. of 43rd Annual Meeting on Association for Computational Linguistics, pp. 141-148, 2005.

[34] M. Lapata, and R, Barzilay, "Automatic Evaluation of Text Coherence: Models and Representations," in Proc. of IJCAI, pp. 1085-1090, 2005.

[35] V. W. Feng, and G. Hirst, "Extending the Entity-based Coherence Model with Multiple Ranks," in Proc. of EACL, pp. 315-324, 2012.

[36] Z. H. Lin, H. T. Ng, and M. Y. Kan, "Automatically Evaluating Text Coherence Using Discourse Relations," in Proc. of ACL, pp. 997-1006, 2011.

[37] 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.

[38] M. Abdolahi, and M. Zahedi, "A New Model for Text Coherence Evaluation Using Statistical Characteristics," Journal of Electrical and Computer Engineering Innovations (JECEL), vol. 6, no. 1, pp. 15-24, 2018.

[39] S. Liu, S. Zeng, S. Li, "Evaluating Text Coherence at Sentence and Paragraph Levels", arXiv preprint arXiv:2006.03221, 2020.

[40] H. P. Luhn, " A business intelligence system," IBM Journal of Research and Development, vol. 2, no. 4, pp. 314-319, 1958.

[41] 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.

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

[43] R. Barzilay, and M. Lapata, "Modeling local coherence: An entity-based approach," Computational Linguistics, vol. 34, no. 1, pp. 1-34, 2008.

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

[45] J. Li, M. T. Luong, and D. Jurafsky, "A Hierarchical Neural Auto encoder for Paragraphs and Documents," in Proc. of ACL, pp. 1106-1115, 2015.

[46] C. Lioma, F. Tarissan, J. Simonsen, C. Simonsen, and B. Larsen, "Exploiting the bipartite structure of entity grids for document coherence and retrieval," in Proc. of the 2016 ACM International Conference on the Theory of Information Retrieval, pp. 11-20, 2016.

[47] 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. of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1087-1096, 2015.

[48] D. T. Nguyen, S. Joty, " A neural local coherence model," in Proc. of the 55th Annual Meeting of the Association for Computational Linguistics., vol. 1, pp. 1320-1330, 2017.

[49] Annual Meeting of the Association for Computational Linguistics, pp. 93-103, 2013.

[50] C. Petersen, and J. G. Simonsen, "Graph-based coherence modeling for assessing readability," in Proc. of the 2015 International Conference on The Theory of Information Retrieval, pp. 191-200, 2015.

[51] M. Mesgar, and M. Strube, "Entropy and graph-based modelling of document coherence using discourse entities: An application to IR," in Proc. of the fourth joint conference on lexical and computational semantics, pp. 309-318, 2015.

[52] H. C. Moon, T. Mohiuddin, S. Joty, X. Chi, "A Unified Neural Coherence Model", arXiv:1909.00349 [cs.CL], 2019.

[53] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in neural information processing systems, pp. 3111-3119, 2013.

[54] M. Abdolahi, and M. Zahedi, "Textual coherence improvement of extractive document summarization using greedy approaches and word2vec," Journal of Modern Education and Computer Science (MECS), vol. 11, no. 4, pp. 23-31, 2019.

[55] M. Moradi, and M. R. Sadeghi, "Combining and Steganography of 3-D Face Textures", Journal of Modern Education and Computer Science (MECS), vol. 5, no. 2, pp. 93-100, 2017.

[56] E. Clark, A. Celikyilmaz, N. A. Smith, "Sentence Mover’s Similarity: Automatic Evaluation for Multi-Sentence Texts", Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2748-2760, 2019.