Text Coherence Analysis based on Misspelling Oblivious Word Embeddings and Deep Neural Network

Md. Anwar Hussen Wadud[1], Md. Rashadul Hasan Rakib[2]
Department of Computer Science and Engineering
Mawlana Bhashani Science and Technology University
Tangail, Bangladesh

***Abstract*—Text coherence analysis is the most challenging task in Natural Language Processing (NLP) than other subfields of NLP, such as text generation, translation, or text summarization. There are many text coherence methods in NLP, most of them are graph-based or entity-based text coherence methods for short text documents. However, for long text documents, the existing methods perform low accuracy results which is the biggest challenge in text coherence analysis in both English and Bengali. This is because existing methods do not consider misspelled words in a sentence and cannot accurately assess text coherence. In this paper, a text coherence analysis method has been proposed based on the Misspelling Oblivious Word Embedding Model (MOEM) and deep neural network. The MOEM model replaces all misspelled words with the correct words and captures the interaction between different sentences by calculating their matches using word embedding. Then, the deep neural network architecture is used to train and test the model. This study examines two different types of datasets, one in Bengali and the other in English, to analyze text consistency based on sentence sequence activities and to evaluate the effectiveness of this model. In the Bengali language dataset, 7121 Bengali text documents have been used where 5696 (80%) documents have been used for training and 1425 (20%) documents for testing. And in the English language dataset, 6000 (80%) documents have been used for training and 1500 (20%) documents for model evaluation out of 7500 text documents. The efficiency of the proposed model is compared with existing text coherence analysis techniques. Experimental results show that the proposed model significantly improves automatic text coherence detection with 98.1% accuracy in English and 89.67% accuracy in Bengali. Finally, comparisons with other existing text coherence models of the proposed model are shown for both English and Bengali datasets.**

***Keywords*—Coherence analysis; deep neural network; distributional representation; misspellings; NLP; word embedding**

I. Introduction

Text coherence analysis is a very well-known key term in natural language processing for a text with multiple sentences [1]. According to Mann and Thompson (1988), a text is coherent in explaining the role that each paragraph plays in the whole field. Text coherence measures the degree of logical consistency for text which is a key property of any well-organized text document. With the rapid development of digital communication mediums such as social networks, mobile devices, or online news portals it is more complex to identify which information is consistent or inconsistent. Recently, paperless assessment has increased rapidly and computers have been used to evaluate assessment. It is very difficult to check the consistency of text among sentences with sort time without automatic evaluation. In social networks or mobile communication, users usually use short text for their communication or use their mobile devices for any type of online assessment. During digital communication or online assessment or reporting news sometimes a naive user may misspell some word or couple of words in their whole text [2]. Common errors such as grammatical mistakes, vocabulary, or syntax errors can easily be determined, but finding text coherence between paragraphs is very difficult both in the manual and computerized systems. It is very important to automatically identify which news or information is valid or coherent regarding other information. The following examples show text coherence and news inconsistency. One example is shown in Table I where Text 1 has logical consistency, but in Text 2, the first sentence and the second sentence are logically coherent but the second sentence with the third sentence is not logically consistent.

Text coherence analysis is very important for many reasons. For example, they can be used as the logical bridge between different words, sentences, and paragraphs. Readers easily detect ideas within each sentence and paragraph. Text or paragraph without coherence not only makes it difficult to determine the main idea but also reads the full text. Text coherence checking in a short portion of the text is very easy for humans, but in a large document that has thousands of paragraphs or more is also difficult and time-consuming.

Dealing with text coherence using a machine is very difficult. Foltz in 1998 [3] proposed the first text coherence evaluation method using a machine. Later, many researchers proposed several text coherence methods, but unfortunately, no method is perfect for finding text coherence between words or sentences or paragraphs. Many automatic summarization methods that can extract summaries from a paragraph can also check grammatically correct but are limited for text coherence analysis. Considering coherence needs to check discourse relations [4], finding common patterns during sentence connection [5].

TABLE I. EXAMPLE OF LOGICAL CONSISTENT AND INCONSISTENT AMONG SENTENCES

Bengali	**English**
জাতীয়দলে খেলতে গিয়ে একজন ফুটবলার তার বাম পা ভেঙে ফেলেন। ডাক্তার তাকে ৩০ দিনের জন্য বিশ্রামের পরামর্শ দিয়েছেন। এজন্যে, তিনি তার নিয়মিত অনুশীলন সাময়িক সময়ের জন্য বন্ধ করেন।	A footballer broke his left leg while playing for the national team. The doctor advised him to rest for 30 days. Because of this, he stopped his regular practice for a while.
Text1: label=1(coherent)	
জাতীয়দলে খেলতে গিয়ে একজন ফুটবলার তার বাম পা ভেঙে ফেলেন। ডাক্তার তাকে ৩০ দিনের জন্য বিশ্রামের পরামর্শ দিয়েছেন। এজন্যে, তিনি খুব তাড়াতাড়ি বিছানা থেকে উঠেন এবং সকালে নিয়মিত অনুশীলন করেন।	A footballer broke his left leg while playing for the national team. The doctor advised him to rest for 30 days. Because of this, he gets out of bed very early and practices regularly in the morning.
Text2: label = 0 (incoherent)	

Recently, proposed coherence analysis methods [1, 6] have been based on a deep learning framework that uses recurrent and recursive neural networks for computing word vectors in sentences. They capture the interaction between sentences by identifying a set of coherence features and computing the similarity between words which is useful for coherence assessment but the main limitation is finding the right word to measure similarity or interaction between sentences. If we identify misspelling sentences and determine word vectors for correct words from a misspelled word, it is a new dimension for coherence analysis. Here is another example of text coherence with misspelling words shown in Table II.

TABLE II. TEXT IN CONSISTENT WITH MISSPELLED WORDS

Bengali	**English**
টম তার চর্বিযুক্ত শরীর নিয়ে খুব অমন্তুষ্ট **(অসন্তুষ্ট)।** তিনি চাবিযুক্ত **(চর্বিযুক্ত)** খাবার খেতে এবং বিয়ার পান করতে পছন্দ করেন তবে কোনও শারীরিক অনুশীলন করেন না। সুতরাং, আমি মনে করি তার নিয়মিত শারীরিক অনুশীলন করা উচিত।	Tom is very dissatified **(dissatisfied)** with his fat body. He likes to eat fatty food and drink beer but has not done any phygical **(physical**) exercise. So, I think she should do regular physical exercise.

The above example contains some misspelled words such as "অমন্তুষ্ট", "চাবিযুক্ত, "dissatified", "phygical" which is logically inconsistent based on existing text coherence analysis methods. The composition is logically consistent between sentences when the correct word is used for each misspelled word such as (অমন্তুষ্ট => অসন্তুষ্ট), (চাবিযুক্ত => চর্বিযুক্ত), (dissatified => dissatisfied), (phygical => physical). Existing text coherence analysis methods convert each word into multidimensional word vectors using pretrained word embedding vectors such as Word2vec [7, 8] and Glove [9] and calculate the text set by considering the semantic and syntactic relationship between sentences. However, they did not work on out of vocabulary or misspelled words, and sometimes their results show that any composition is locally inconsistent between the sentences that are logically coherent.

Finding correct word from misspelling word is very challenging work both in Bengali and English language processing task. In Bengali language there have lot of variation in word formation. Changing single character in a word can modify the meaning of a single sentences. In this paper, a modern text coherence analysis method using Misspelling Oblivious Word Embedding Model (MOEM) and deep neural network has been proposed to overcome the above limitation. A set of commonly misspelled words is identified with their correct words to find similarities between sentences and study the coherence problem with a set of coherence features. First, misspelling oblivious word embedding methods generate a sentence matrix with the correct word vector and then apply a deep neural network with a set to cohere to compute the similarity among sentences. Finally, the proposed method estimated the text coherence by combining word vectors and similarity scores. The main contributions of this study are as follows:

- Develop a corpus of 12000 text documents for English and 8000 text documents for Bengali with misspelling words from different social media and newspaper;
- Label each document as coherent and incoherent after performing cleaning, stemming, stop-words removal, normalization and tokenization;
- Identify misspellings with the correct spelling using the MOEM model from the misspelled word dictionary set.
- Design a coherence model to identify English and Bengali documents into coherent and incoherent categories.
- Compare the performance of the proposed text coherence model with the existing models.

The rest of the paper is organized as follows: Section 2 introduces a review of recent work in this field; A detailed explanation about the proposed model is presented in Section 3. Section 3 describes the development of a data corpus for Bengali and English with misspelled word models and calculates their word metrics; Section 4 discusses the experimental setup and performance analysis results of the proposed model; Finally, Section 5 concludes the paper and highlights the importance of text analysis in both English and Bengali, including summaries and future opportunities.

II. Related Work

In this section, the main categories of existing text coherence analysis methods are reviewed and described. In 1998 Flotz [4] proposed the first text coherence evaluation model. In his model, text coherence is defined by checking semantic relatedness between sentences that are adjacent to each other where lexical meaning is used to compute semantic relatedness which is a vector-based representation. Since 1998, many researchers proposed several text coherence analysis models such as entity-based model [10-16], syntactic pattern-based models 17], discourse relation-based models [18], content-based model via Hidden Markov Model [19, 20], coreference resolution-based model [21, 22] and cohesion-driven based model [23]. These models use a supervised learning approach to obtain text coherence by computing the relationship between adjacent sentences based on the lexical chain [13, 24] which is the lexical cohesion structure representation of a text.

The entity-based text coherence analysis model is one of the most popular methods that analyses the grammatical role of words in adjacent sentences and evaluates the local coherence [25] by extracting a pattern from adjacent sentences. Initially, R. Barzilay, M. Lapata [12, 14, 26] proposed the model but in recent years some modern approaches such as neural network models [27] and original bipartite graph [28] models, were proposed to overcome the limitation ability of entity grade to detect consistency in just neighbor sentences [29].

Petersen and Simonsen [30] proposed another novel method based on graph theory and the entropy method for measuring the consistency between sentences in a document. In their model they increased more nouns in the document which increases adverse information in text focusing and is the limitation of the lower score for global coherence analysis. Another graph theory-based novel model was introduced by M. Mesgar [31] for coherence features based on frequent subgraphs where texts are consistent with particular patterns in extracted subgraphs and compare their ability to measure the readability of Wall Street Journal articles [31] using an entity graph coherence model.

Another popular text coherence evaluation [32-34] approach is based on statistical machine translation algorithms such as the EM and IBM Algorithms [12, 24], where the meaning of each word in the target language represents several words and each word establishes a link into multiple sentences and finds coherence using this link word. However, these algorithms cannot overcome semantic feature limitations. Modern approaches such as neural networks [3], deep neural models [1, 3], recurrent neural networks (RNNs) [35, 36], etc. overcome semantic feature limitations and sentence ordering problems by using distributed representation and extracting syntactic representation of discourse coherence [3]. A deep neural network tries to calculate local and global coherence [37] and the RNN network is used to obtain distributed representation [3] of the sentences and sequence modeling tasks [34]. Sennan [42] proposed a text coherence model based on the sentence ordering task, but they did not discuss words outside of vocabulary or misspellings. Nevertheless, there are some limitations because text coherence analysis is a very challenging task in natural language processing. Different from the above studies, a new text coherence model has been proposed where the deep neural network is to find sentence discourse coherence and Misspelling Oblivious Embeddings [2] used to obtain the correct format and actual meaning of several words in a document.

III. Proposed Model

The main goal of this work is to develop an architecture based on a deep learning network that can predict text coherence in both Bangla and English text documents. Fig. 1 shows a simple process of the proposed system consisting of four main phases: preprocessing, feature extraction, training, and prediction. The following subsections outline detailed explanations for each level of the proposed system.

A. Collecting Data and Preprocessing

The two most widely used corpora [12, 17, 20, 22, 41], one is a collection of aviation accident reports, and the other corpus is American earthquake-related news has been used for English text coherence analysis. The accident reports-related dataset contains 4500 compositions, where per composition have 11 sentences on average and American earthquake-related news has 3000 compositions, with an average of 10 sentences in each composition.

However, for the Bengali language, no dataset is available to identify the textual consistency of any text document. In this study, data was crawled from several social media, online newspapers, etc. A total of 7121 text documents have been crawled where 3565 texts are consistent and 3556 texts are in inconsistent class. Crowd-sourced data are initially labeled according to coherent and incoherent class. Table III summarizes some of the features of the developed dataset.

Preprocessing is used to convert raw data into a state where the machine can easily parse it. Several techniques were used for data preprocessing. The tokenization technique is used to convert data from sentences to words by dividing the sentence into sets of tokens. The text clear and stop word removal technique is used to remove special characters, punctuation, numbers and unnecessary words. In this research The Natural Language Toolkit (NLTK) is used to complete preprocessing where NLTK provides all text processing libraries, such as tokenization, stemming, parsing, tagging, and semantics reasoning.

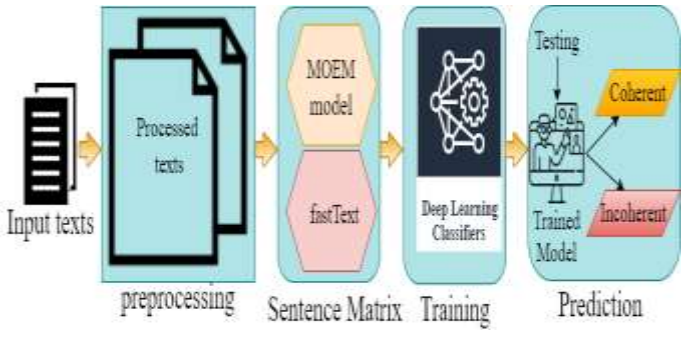

Fig. 1. Simple Process of the Proposed Text Coherence Detection System.

TABLE III. DATASET SUMMARY

Properties	Bengali Data Corpus		English Data Corpus	
	Coherent	*Incoherent*	*Coherent*	*Incoherent*
Total documents	3565	3556	3780	3720
Number of words	122345	233558	148388	225952
Unique words	15450	18490	16735	21687
Avg. words per doc.	34.32	65.68	39.25	60.74
Max. text length	340	2310	510	2690
Min. text length	4	10	1	5
Number of misspelling words	1500	2500	1200	2000

B. Sentence Matrix

Each sentence contains a combination of several meaningful words that must be translated into true-value feature vectors, and the combination of all vectors is used to form a sentence matrix. Some words in a sentence may be misspelled words or even out of vocabulary that cannot be directly translated into feature vectors. The misspellings embedding (MOE) [2] model is used to find the correct word from misspelled words.

C. Word Embedding with Misspelling Word Model

Facebook introduces a new word embedding method named Misspelling Oblivious Embeddings (MOE) [2] which extends fastText [38] architecture to handle out-of-vocabulary (OOV) [2] limitations during natural language processing. fastText was built by extending Word2Vec architecture which uses skip-gram models with negative sampling and the SoftMax activation function.

Popular pretrained word embedding methods such as word2vec [7, 8], GloVe [9], fastText [38], etc. provide word vectors during training but fail to produce word embedding when words are out-of-vocabulary (OOV). MOE word embedding methods work by considering slang, misspellings, or abbreviations. The MOE calculated the weighted sum of two-loss functions which are the semantic loss and spell correction loss functions. The semantic loss function captures the semantic relationship between words denoted by L_{FT} and spell correction loss function map words to find the correct word embedding denoted by L_{SC}. The MOE [2] is defined as follows:

$$L_{MOE} := (1 - \alpha)L_{FT} + \alpha \frac{|T|}{|M|} L_{SC} \quad (1)$$

where α is the hyperparameter, T is the text corpus and M is the misspellings dataset. Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract.

The misspelling model has a large vocabulary [2] and a set of pairs (misspelling, correction) for the spell correction word. This network model is used to process the first part of our datasets, where there are misspelled words, out-of-vocabulary words, etc. Then we apply the fastText [38] model to obtain word embedding vectors of our corpus.

Fig. 2 shows the word embedding generation process using a fastText model where the N-gram method splits the word into subwords. For example, “orange” word can be split into “ora”, “nge”, “ang”, “oran”, “rang”, “ange”, “orang” and “range” subwords, and the sum of all subword embedding vector is considered as the embedding of “orange” word.

Similarly, Fig. 3 shows the misspelled word embedding generation process where each word consists of pairs (X, C) of a word where X denotes all possible misspelled words as shown in Table IV for a specific correct root word C. If a word is combination of two, three or more root words as shown in Table V then MOE [2] use N-gram method to split the word into all root words then find the correct root word from misspelled word and calculates word embedding by performing doc products between the sum of input vectors of the misspelled word and correct word.

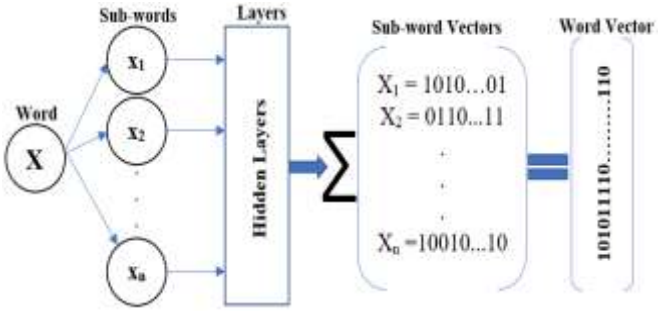

Fig. 2. Generating Word Embedding using FastText.

TABLE IV. SOME ROOT WORD WITH MISSPELLED WORD

Misspelled Subword(X)	Correct Spelled Subword(C)
অমন্তষ্ট, অমুন্ষ্ট, আসন্তষ্ট, অসিন্তষ্ট, অসান্ষ্ট, etc.	অসন্তষ্ট
দুরি, দুরই, ধুর, দরু, দারু, etc.	দূর
ছাকরি, চাকিরি, চাখরি, চাকরী, etc.	চাকরি

TABLE V. SOME COMBINE WORD WITH ROOT WORD

Combine Word	Root words	Combine Word	Root words
বাসস্ট্যান্ড	বাস, স্ট্যান্ড	নিশীথবনবিলাসিনী	নিশীথ, বন, বিলাসিনী
আইনজীবী	আইন, জীবী	অঘটনঘটনপটিয়সী	অঘটন, ঘটন, পটিয়সী
পূর্ণমান	পূর্ণ, মান	উষ্ট্রকন্টকভোজননন্যায়	উষ্ট্র, কন্টক, ভোজন, ন্যায়

D. Proposed Coherence Detection Architecture

Coherence can be detected by considering all text or paragraphs in a composition or considering two consecutive paragraphs. However, there can be another way to find text coherence that considering any two or three paragraphs makes the whole article semantically coherent which is applied in our model. For example, if one chooses the first sentence and last sentence from multiple sentences in composition and finds coherence then it will be said that the composition is semantically consistent.

1) Model inputs: Since this study considers words out of vocabulary, misspelled words, etc. therefore, the input of the proposed coherence model will be the output of the misspelling word embedding model which are word vectors of different types of words. Every time the proposed architecture considers three paragraphs as input into the model from the composition. Then determine the word vector of the selected paragraph and iterate the process until all paragraphs are selected.

2) Proposed text coherence methods: A deep learning [1,39,40] network is used to process the current word embedding output and train the model based on a large Bengali and English data corpus. Fig. 4 shows the proposed model where each time processes three paragraphs to find text coherence among three sentences. A word embedding matrix with a 50-dimensional size is formed by concatenating all word vectors in a finite size vocabulary. Convolutional neural networks have been applied to the proposed model and various filters have been used which is a matrix of weights to extract useful patterns from input sentences. The global max pooling layer and rectified linear (ReLU) function defined as max (0, x), are used to increase the accuracy after the convolution layer. Then, the first sentence is concatenated with the second sentence and second sentence with the third sentence and third sentence is concatenated with the first sentence to compute sentence-to-sentence similarity. This model used several hidden and softmax layers which are a series of convolutional and pooling operations, to find coherence the probability of these three sentences.

3) Prediction: Sigmoid activation function is used to calculate coherence probabilities of three sentence in output layer. This activation function also used as threshold on testing set. The trained classifier model has been used for coherence prediction based on testing set. If the threshold is Th and predicted probability is P then predicted class C can be defined as:

$$C = \begin{cases} Incoherent, \ if \ P \leq \mathrm{T}_h \\ Coherent, \ if \ P > \mathrm{T}_h \end{cases} \quad (2)$$

Since the proposed text coherence model classifies coherent and incoherent classes as binary classifications, the sigmoid activation function is used without changing the default threshold value.

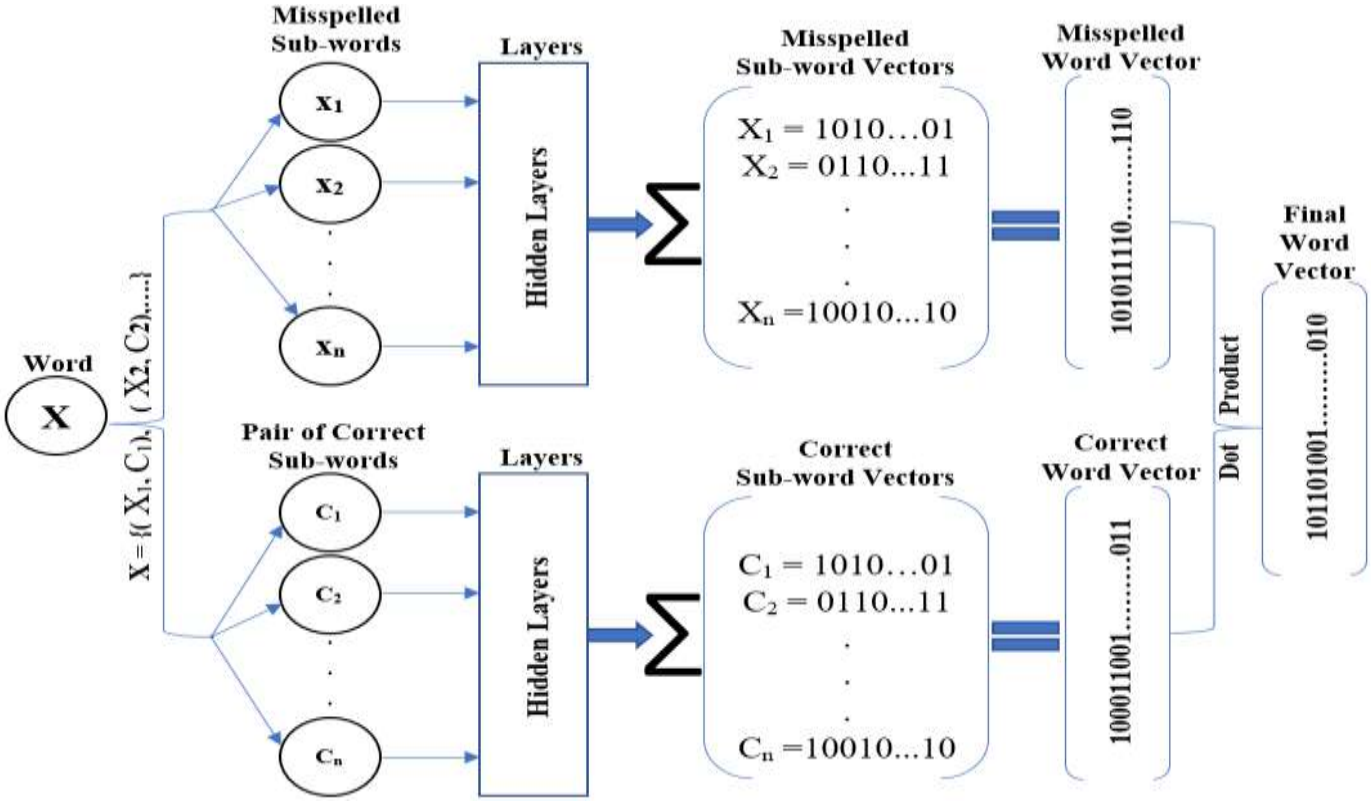

Fig. 3. Misspelling Word Embedding Generation Process using FastText.

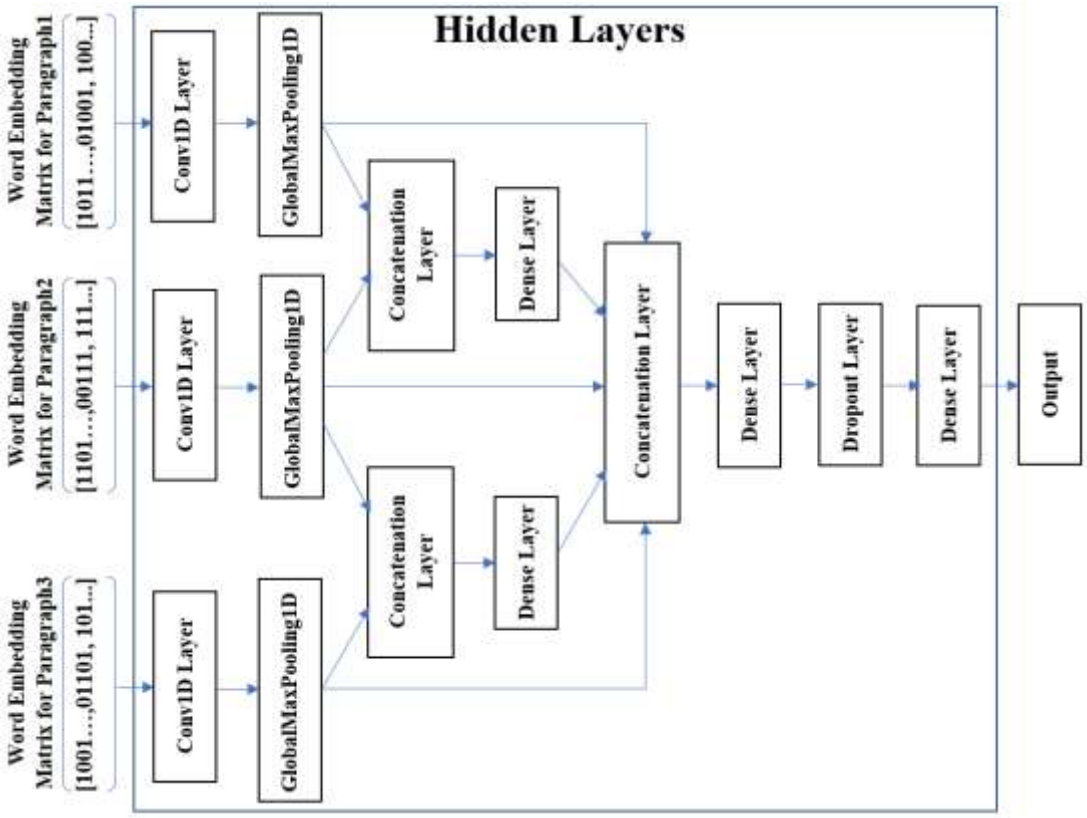

Fig. 4. Proposed Text Coherence Model.

IV. Experimental Setup

The goal of the experiments is to evaluate the proposed model for two types of datasets and make a comparison between the proposed model and other coherent analysis models. The open-source Google Colab platform was used to conduct the test experiment where the Python version was 3.7, TensorFlow was 2.2.1, Pandas 1.3.3, and Scikit-learn version was 0.22.2. Panda data frame is used for data set preparation and Scikit-learn for training and testing purposes. This study initially examined the dataset from each original document by setting the dimension size of the matrix to 50, the size of the convolution filter to 4, the batch size to 500, and the total number of epochs to 20. Datasets for testing and training, the proposed model uses 80% of the total data for the train and 20% of the total dataset as the test dataset. Every dataset contains misspelled words, out-of-vocabulary words, punctuation, etc. fastText and MOE model used to compute the word embedding matrix for each data corpus. fastText pretrained Bengali word embedding vectors and MOE models are used for misspelling words to construct a word embedding matrix for Bengali text documents. A collection of pairs (misspellings, corrections) with 2,746,061 vocabulary sizes has been used for Bengali and English language datasets to obtain the proper word vector of the misspelled word or out of the vocabulary word in a data corpus. For both the Bengali and English lingual datasets, positive samples are labeled with 1, and 0 is labeled for all negative samples.

A. Measures of Evaluation

Statistical and graphical measure are used to show the performance of the proposed system based on Accuracy, Precision, Recall and F1-score.

1) Accuracy: A mathematical measure that indicates that a classifier correctly classifies or prohibits a condition. This is known as the symmetry of the actual results in the amount of samples tested.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Where, TP = True positive; FP = False positive; TN = True negative and FN = False negative.

2) Precision: is the ratio of how many text documents are actually coherent among the whole documents. Precision is defined by.

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

3) Recall: is the ratio of how many text documents are classified correctly as coherent among total coherent text.

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

4) F1-Score: The weighted average of accuracy and precision. This mathematical assessment metric is used to decide which of these different classifications needs to be chosen.

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

B. Result of Proposed Model on English Data Corpus

For the training and evaluation of the English data corpus, positive cliques have been used as coherent documents from the original training document, and other documents contain sentences that were replaced by each other in a set of negative clique datasets. The proposed model has been applied on the English data corpora and made a comparison of proposed results with other existing methods such as DCM [1], Recursive [3], Recurrent [3], Entity Grid [14], HMM [17], HMM + Content [17], Conference + Syntax [14], and Graph [29] as shown in Table VI. According to Table VI proposed model achieves better performance than all other existing coherent frameworks.

TABLE VI. COMPARISON OF DIFFERENT COHERENCE MODELS ON ENGLISH DATA CORPUS

Model Name	Accident	Earthquake	Average
Proposed Model	0.986	0.977	0.981
DCM	0.950	0.995	0.973
Recursive	0.864	0.976	0.920
Recurrent	0.840	0.951	0.895
Entity Grid	0.904	0.872	0.888
HMM	0.822	0.938	0.880
HMM + Content	0.742	0.953	0.848
Conference + Syntax	0.765	0.888	0.827
Graph	0.846	0.635	0.741

Compared with the DCM model, proposed model generates a strong semantic relationship between sentences by using the misspelling oblivious word embedding model which is missing in other text coherence analysis methods. The deep coherence model uses a convolutional neural network for text coherence assessment and word2vec as pretrained word embedding vectors for matrix construction of each sentence so that out of the context word, it cannot calculate and sometimes constructs an incorrect sentence matrix. Proposed model used fastText as pretrained word embedding vectors to compute the word embedding for sentence matrix construction and used the misspelling oblivious word embedding model to calculate out of vocabulary words and produced a better result than the DCM model.

HMM and Entity Grid require manual feature engineering and sentence representation where the proposed model can automatically learn sentence representation. The recursive and recurrent models use syntactic parsers to construct a syntactic tree and then calculate semantic coherence, which requires expensive preprocessing time. Proposed model uses a deep neural network for automatic preprocessing and makes the effort required of feature engineering unnecessary.

Fig. 5 shows the text coherence analysis accuracy result as a pie diagram of accidental data corpus and the accuracy of the text coherence on the earthquake data is shown in Fig. 6. In Fig. 5, proposed model produces 13% accuracy which is higher than other coherence models on accidental data in the English language. In the accidental data corpus, there are some misspelled words so existing models cannot evaluate these error words. As a result, their accuracy score is lower than the proposed model. However, the proposed model shows 12% accuracy in Fig. 6 which is equal to the accuracy of other models named DCM, recursive, and HMM + content text. Because there is no misspelled word in the test data set. So, the proposed model produces equal accuracy likes other coherence models.

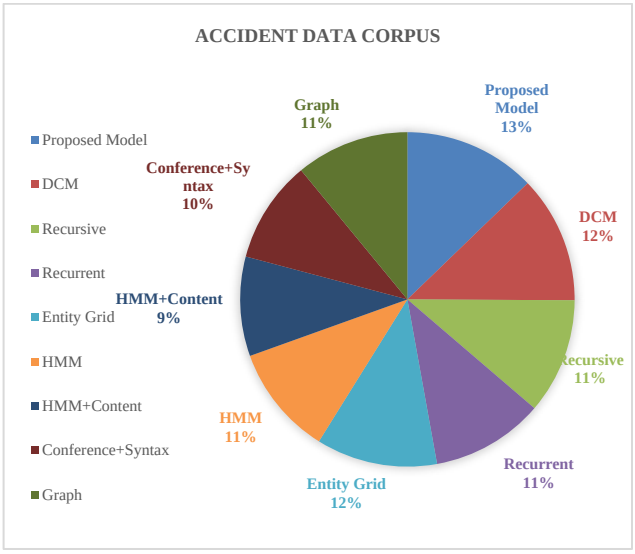

Fig. 5. English Text Coherence Analysis on Accidental Data.

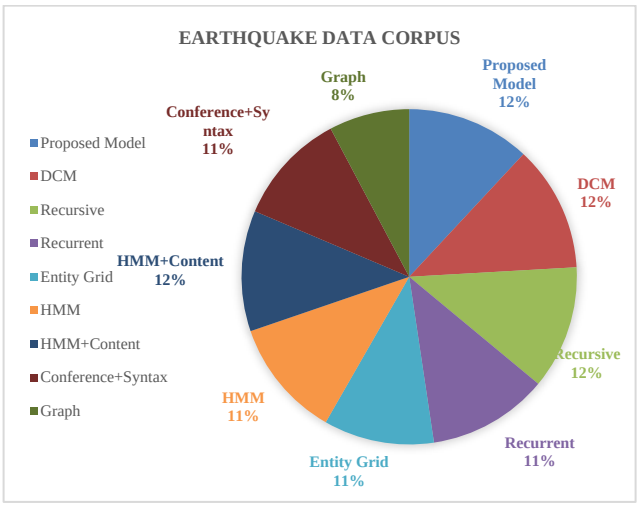

Fig. 6. English Text Coherence Analysis on Earthquake Data.

C. Result of Proposed Model on Bengali Data Corpus

There is no single standard method for analyzing Bengali text consistency. Five separate classification algorithms are used to evaluate the proposed system to find the best method for analyzing Bangla text continuity. To calculate for the best accuracy, the first Bengali dataset was created without considering misspellings and applied the experiment of all text coherence models to the datasets. Table VII reports the results of the proposed model and a comparison of other existing text coherence methods, such as DCM [1], Recursive [3], entity

grid [14], and HMM [17], for the Bengali data corpus. The proposed model has achieved maximum accuracy of 80.46% where the maximum precision value from HMM model is 85.30% and the maximum recall value of 95.87% obtained from the fastText model.

Table VIII shows a comparison of performance between different text coherence models on misspelled words. First, the fast text model has been applied to the Bengali dataset, but the training datasets have a lot of out of vocabulary and misspellings words. Fast text word embedding vectors cannot generate an actual sentence matrix for all Bengali words, and text consistency accuracy is very low in Bengali. Then the proposed model with MOE method has been applied in Bangla data corpus and achieved better results than the fast text model. Using the MOE method, significant changes have been made and more accurate accuracy has been shown for the Bangla data corpus. Other text integrated methods, such as DCM, Entity Grid and HMM models, are applied to the Bangla data corpus and produce lower results than the proposed model.

Fig. 7 depicts the f1-score of different text coherence technique without considering misspelled words where proposed model achieved maximum of 83.38% f1-score and lowest f1-score is 20.37% obtain from HMM text coherence method.

Similarly, Fig. 8 shows the f1 scores of various coherent models applied to the dataset containing misspelled words. The F1 score suggests that the proposed model is more suitable for text consistency analysis than other existing models. This is because the proposed model has achieved the highest F1 score for both coherent (90.06%) and inconsistent (88.57%) classes.

TABLE VII. COMPARISON OF DIFFERENT COHERENCE MODELS ON BENGALI DATA CORPUS WITHOUT CONSIDERING MISSPELLEING WORDS

Model Name	Accuracy (%)	Precision (%)	Recall (%)
Proposed model	80.46	76.29	91.90
fastText Model	79.40	73.66	95.87
DCM	78.67	78.99	80.52
Recursive	73.37	68.41	92.87
Entity Grid	62.47	59.83	89.84
HMM	56.46	85.30	21.54

TABLE VIII. COMPARISON OF DIFFERENT COHERENCE MODELS ON MISSPELLEING WORDS

Model Name	Accuracy (%)	Precision (%)	Recall (%)
Proposed model	89.67	89.68	90.46
fastText Model	78.94	74.28	93.66
DCM	72.24	67.14	93.24
Recursive	75.82	73.84	84.42
Entity Grid	64.08	77.36	45.83
HMM	52.39	93.69	11.43

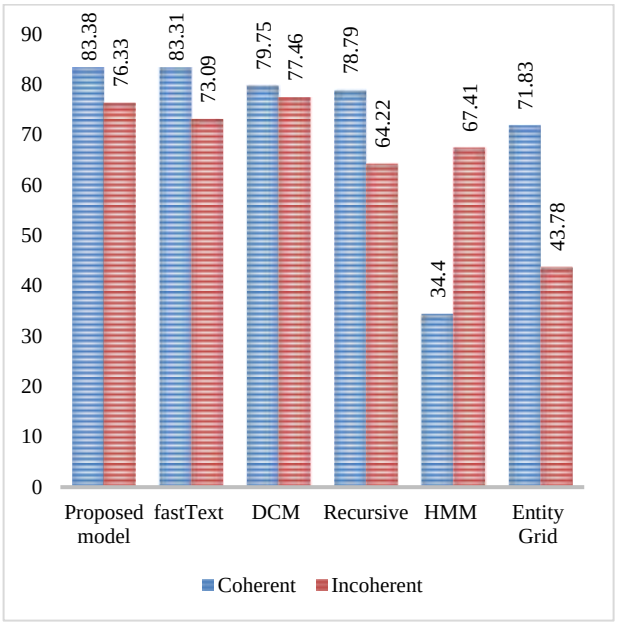

Fig. 7. F1-Score of different Text Coherence Model without Considering Misspelling Words.

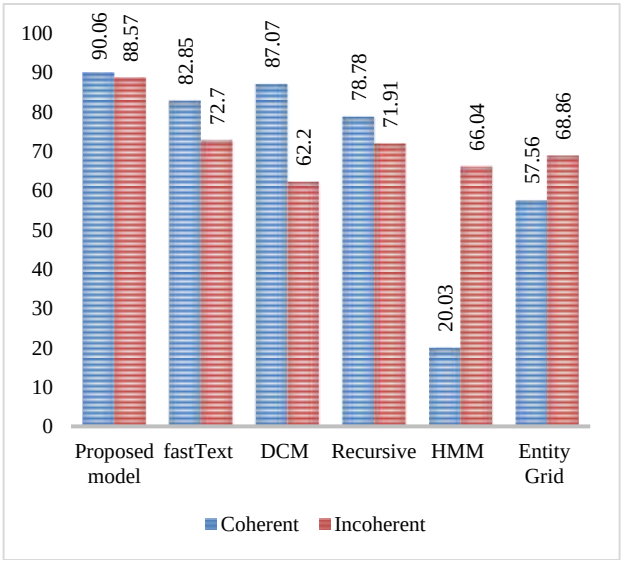

Fig. 8. F1-Score of different Text Coherence Model with Misspelling Words.

The Receiver Operator Characterization (ROC) curve is an important evaluation metric that plots the probability of True Positive Rate (TPR) as opposed to the False Positive Rate (FPR) of different Threshold values and shows the Area Under the Curve (AUC) of various machine learning classifiers. The ROC curve analysis of the various text coherence models shown in Fig. 9 and 10. The ROC curve in Fig. 9 is drawn from a general data set where misspelled words are not considered here. Proposed model obtained the maximum AUC value of 79.7% where the AUC value of other text coherence model is lower than the proposed model. The Fast Text model and the DCM model provide similar AUC values of 78% but HHM Text coherence model shows 58.70% AUC values which is too poor for text coherence classification.

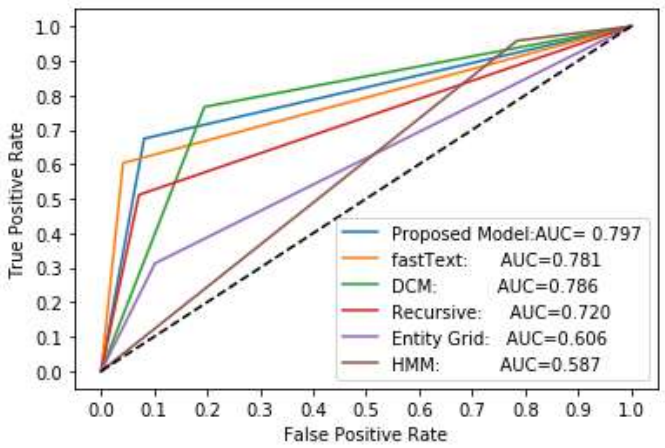

Fig. 9. ROC Curve of different Text Coherence Model without Considering Misspelling Words.

The proposed method gives better results when misspelled words are considered during the test and AUC value is 89.30% as shown in Fig. 10 where accuracy is much higher than the previous AUC value of 79.70% shown in Fig. 9. However, the AUC value of other text-based models remains the same as the previous AUC values. The AOC values of the existing model in Fig. 10 are slightly different but not like the proposed model. This presents that the proposed model performs more accurately during the classification of coherence classes in the dataset.

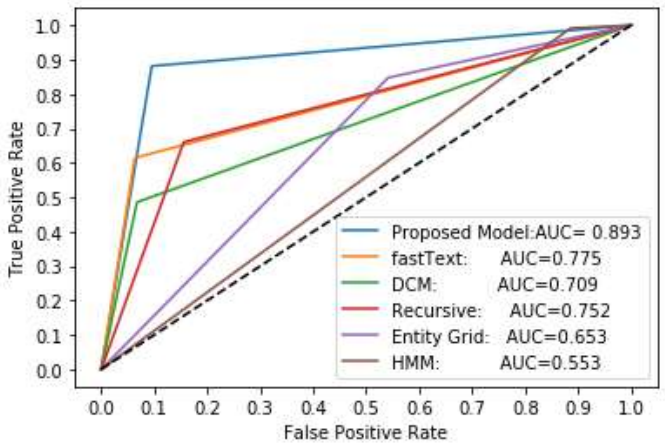

Fig. 10. ROC Curve of different Text Coherence Model with Misspelling Words.

The test data contains some misspelled words and the proposed model gives better results than other text coherence models because other methods may not work on misspelled words which is the main imitation of all other methods. This presents that the proposed model performs more accurately during the classification of coherence classes in the dataset.

V. Conclusion

In this research paper, the main objective of this study was to calculate the text consistency with misspelled words in Bengali language. A model has been proposed based on a deep neural network and MOE method for text coherence analysis. For experimental analysis, both Bengali and English data corpora have been tested and the proposed model shows significant improvement in text coherent assessment. The proposed model shows an average accuracy of 98.1% in English text coherence analysis for datasets considering misspelled words which is higher than the existing models. For the analysis of Bengali text coherent, this study experimented on two types of datasets. One type of dataset contains common words and another type of dataset contains out-of-vocabulary, misspelled words, etc. The proposed model prediction for general datasets shows 80.46% accuracy and misspelled datasets 89.67%. The accuracy of other models is appropriate for general datasets but the accuracy goes down for misspelling datasets. The accuracy of fastText, DCM, Recursive, Entity Grid, and HMM models is 79.40%, 78.67%, 73.37%, 62.47%, and 56.46% respectively for normal dataset but for misspelling dataset accuracy is 78.94%, 72.24%, 75.82%, 64.08% and 52.39% respectively which is less than the normal dataset accuracy. However, the proposed model performs better accuracy for both normal and misspelled datasets and increases the accuracy for misspelling dataset than a normal dataset. Currently, this study uses limited (key, value) pairs for word misspelling but for more accuracy, it requires a huge collection of word pairs for misspelled words which is the main limitation of the proposed model.

References

[1] B. Cui, Y. Li, Y. Zhang and Z. Zhang, Text coherence analysis based on deep neural network, CoRR abs/1710.07770(2017).

[2] B. Edizel, A. Piktus, P. Bojanowski, R. Ferreira, E. Grave and F. Silvestri, Misspelling oblivious word embeddings, CoRR abs/1905.09755(2019).

[3] J. Li and E. Hovy, A model of coherence based on distributed sentence representation01 2014, pp. 2039–2048.

[4] P.W.Foltz, W.Kintsch and T.K.Landauer, The measurement of textual coherence with latent semantic analysis, Discourse Processes 25(23) (1998) 285–307.

[5] D.Marcu, The rhetorical parsing of natural language texts, in Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics ACL '98/EACL '98, (Association for Computational Linguistics, USA, 1997), p. 96–103.

[6] H. Oufaida, P. Blache and O. Nouali, A Coherence Model for Sentence Ordering 062019, pp. 261–273.

[7] T. Mikolov, K. Chen, G. Corrado and J. Dean, Efficient estimation of word representations in vector space, in 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings, eds. Y. Bengio and Y. LeCun2013.

[8] T. Mikolov, I. Sutskever, K. Chen, G. Cor-rado and J. Dean, Distributed representations of words and phrases and their compositionality, CoRR abs/1310.4546(2013).

[9] J. Pennington, R. Socher and C. Manning, Glove: Global vectors for word representation, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Association for Computational Linguistics, Doha, Qatar, October 2014), pp. 1532–1543.

[10] D. Xiong, M. Zhang and X. Wang, Topic based coherence modeling for statistical machine translation, IEEE/ACM Transactions on Audio, Speech, and Language Processing 23(3) (2015) 483–493.

[11] C. Petersen, C. Lioma, J. Simonsen and B. Larsen, Entropy and graph based modelling of document coherence using discourse entities: An application (07 2015).

[12] R. Barzilay and M. Lapata, Modeling local coherence: An entity-based approach, in Proceedings of the 43rd Annual Meeting of the Association

for Computational Linguistics (ACL'05) (Association for Computational Linguistics, Ann Arbor, Michigan, June2005), pp. 141–148.

[13] S. Somasundaran, J. Burstein and M. Chodorow, Lexical chaining for measuring discourse coherence quality in test-taker essays, in Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers (Dublin City University and Association for Computational Linguistics, Dublin, Ireland, August 2014), pp. 950–961.

[14] R. Barzilay and M. Lapata, Modeling local coherence: An entity-based approach, Computational Linguistics 34(1) (2008) 1–34.

[15] V. W. Feng and G. Hirst, Extending the entity-based coherence model with multiple ranks, in Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (Association for Computational Linguistics, Avignon, France, April 2012), pp. 315–324.

[16] M. Zhang, V. W. Feng, B. Qin, G. Hirst, T. Liu and J. Huang, Encoding world knowledge in the evaluation of local coherence, in Proceedingsofthe2015ConferenceoftheNorthAmerican Chapter of the Association for Computational Linguistics: Human Language Technologies (Association for Computational Linguistics, Denver, Colorado, May–June 2015), pp. 1087–1096.

[17] A. Louis and A. Nenkova, A coherence model based on syntactic patterns, in Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (Association for Computational Linguistics, Jeju Island, Korea, July 2012), pp. 1157–1168.

[18] Z.Lin, H.T.NgandM.-Y.Kan, Automatically evaluating text coherence using discourse relations, in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (Association for Computational Linguistics, Portland,Oregon,USA,June2011),pp.997–1006.

[19] M. Elsner, J. Austerweil and E. Charniak, A unified local and global model for discourse coherence, in Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference (Association for Computational Linguistics, Rochester, New York, April 2007), pp. 436–443.

[20] R. Barzilay and L. Lee, Catching the drift: Probabilistic content models, with applications to generation and summarization, in Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004 (Association for Computational Linguistics, Boston, Massachusetts, USA, May 2 - May 7 2004), pp. 113–120.

[21] R. Iida and T. Tokunaga, A metric for evaluating discourse coherence based on coreference resolution, in Proceedings of COLING 2012: Posters (The COLING 2012 Organizing Committee, Mumbai, India, December 2012), pp. 483–494.

[22] M. Elsner and E. Charniak, Coreference inspired coherence modeling.01 2008, pp. 41– 44.

[23] Z. G. W. M. XU Fan1, ZHU Qiaoming2, Cohesion-driven discourse coherence modeling,28(3) (2014) p. 11.

[24] D. Xiong, Y. Ding, M. Zhang and C. L. Tan, Lexical chain based cohesion models for document-level statistical machine translation, in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics, Seattle, Washington, USA, October2013), pp. 1563–1573.

[25] K. Filippova and M. Strube, Extending the entity-grid coherence model to semantically related entities, in Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07) (DFKI GmbH, Saarbrucken, Germany, June 2007), pp. 139– 142.

[26] M. Lapata and R. Barzilay, Automatic evaluation of text coherence: Models and representations. Proceedings of the 19th International Joint Conference on Artificial Intelligence 01 2005, pp. 1085–1090.

[27] C. Lioma, F. Tarissan, J. G. Simonsen, C. Petersen and B. Larsen, Exploiting the bipartite structure of entity grids for document coherence and retrieval, CoRR abs/1608.00758 (2016).

[28] F. Xu, S. Du, M. Li and W. Mingwen, An entity-driven recursive neural network model for Chinese discourse coherence modeling, International Journal of Artificial Intelligence Applications 8(03 2017) 1–9.

[29] C. Guinaudeau and M. Strube, Graph-based local coherence modeling ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference 108 2013, pp. 93–103.

[30] C. Petersen, C. Lioma, J. Simonsen and B. Larsen, Entropy and graph-based modelling of document coherence using discourse entities: An application (07 2015).

[31] M. Mesgar and M. Strube, Graph-based coherence modeling for assessing readability, in Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (Association for Computational Linguistics, Denver, Colorado, June 2015), pp. 309–318.

[32] M. Abdolahi and M. Zahedi, A new model for text coherence evaluation using statistical characteristics, Journal of Electrical and Computer Engineering Innovations (JECEI) 6(1) (2018) 15–24.

[33] R. Soricut and D. Marcu, Discourse generation using utility-trained coherence models, in Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions (Association for Computational Linguistics, Sydney, Australia, July 2006), pp. 803–810.

[34] R. Rosenfeld, A maximum entropy approach to adaptive statistical language modeling, Computer, Speech and Language 10 (1996) 187–228.

[35] Y. Pang, J. Liu, J. Zhou and K. Zhang, Paragraph Coherence Detection Model Based on Recurrent Neural Networks 07 2019, pp. 122– 131.

[36] L. Logeswaran, H. Lee and D. R. Radev, Sentence ordering using recurrent neural networks, CoRR abs/1611.02654(2016).

[37] W. Liang, R. Feng, X. Liu, Y. Li and X. Zhang, Gltm: A global and local word embedding based topic model for short texts, IEEE Access 6(2018) 43612–43621.

[38] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. J´egou and T. Mikolov, Fasttext.zip: Compressing text classification models, CoRR abs/1612.03651(2016).

[39] J. Devlin, M. Chang, K. Lee and K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805(2018).

[40] R. Collobert and J. Weston, A unified architecture for natural language processing: Deep neural networks with multitask learning, in Proceedings of the 25th International Conference on Machine Learning ICML '08, (Association for Computing Machinery, New York, NY, USA, 2008), p. 160–167.

[41] M. Elsner and E. Charniak, Extending the entity grid with entity-specific features, in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (Association for Computational Linguistics, Portland, Oregon, USA, June 2011), pp. 125–129.

[42] Liu, Sennan, Shuang Zeng, and Sujian Li. "Evaluating Text Coherence at Sentence and Paragraph Levels." arXiv preprint arXiv:2006.03221 (2020).