FFCD: A Fast-and-Frugal Coherence Detection Method

SWAGATA DUARI AND VASUDHA BHATNAGAR
Department of Computer Science, University of Delhi, New Delhi 110007, India
Corresponding author: Swagata Duari (sduari@cs.du.ac.in)

ABSTRACT In the era of heavy emphasis on deep neural architectures with well-proven and impressive competencies albeit with massive carbon footprints, we present a simple and inexpensive solution founded on Bidirectional Encoder Representations from Transformers Next Sentence Prediction (BERT NSP) task to localize sequential discourse coherence in a text. We propose Fast and Frugal Coherence Detection (FFCD) method, which is an effective tool for the author to incarcerate regions of weak coherence at the sentence level and reveal the extent of overall coherence of the document in near real-time. We leverage the pre-trained BERT NSP model for sequential coherence detection at sentence-to-sentence transition and evaluate the performance of the proposed FFCD approach on coherence detection tasks using publicly available datasets. The mixed performance of our solution compared to state-of-the-art methods for coherence detection invigorates efforts to design explainable and inexpensive solutions downstream of the existing upscale language models.

INDEX TERMS Coherence analysis, discourse analysis, fast-and-frugal coherence detection, writing quality.

I. INTRODUCTION
Computational discourse coherence analysis (CDCA) has gained significant research interest due to recent impetus on NLG models, with summarization and question-answering as significant downstream tasks. Earlier studies on discourse analysis were aligned to linguistics point of view and focused on sentence level analysis [1]–[4]. Later, computational discourse analysis approaches were proposed that build on theoretical developments on coherence analysis [5]–[8]. Recent works in this area are largely based on deep neural approaches that learn to identify latent information and patterns from the text to gauge its extent of coherence [9]–[15].

Deep neural models for CDCA have resulted into a dramatic leap in performance. However, despite their strong ability to perform this pragmatic task, the current approaches have two limitations.

(i) The training and evaluation tasks do not always translate to an equivalent real-life use-case. The adaptability issues listed below pose hindrance in straight-forward application of these models to localize weak semantic connectivity between adjacent sentences, and assist the author to improve writing.

(ii) Often trained on manually annotated large data sets, neural models are typically tested on the same domain/collection. Recent investigations of the true ability of pre-trained language models to detect document coherence in cross-domain setting confirm drop in the model performance to different extents [15], [20]. Further, the expense involved in manual annotation for data-set preparation limits creation of new data-sets resulting into

The associate editor coordinating the review of this manuscript and approving it for publication was Shun-Feng Su.
concerted efforts to improve the benchmarks by experimenting with deep-neural architectures, fine-tuning hyper-parameters, etc. Peyrard [21] candidly informs a similar predicament in text summarization research, where emphasis on empirical approaches has limited furthering the frontiers of technology.

In this study, we investigate the capability of BERT Next Sentence Prediction (NSP) task [22] for analysing discourse coherence. Interpreting BERT NSP score as a cue to coherence in consecutive sentence pairs, we propose a quick and inexpensive method to quantitatively assess the quality of written discourse. We first isolate incoherence by localizing points of extremely weak coherence (unusually low NSP scores). Subsequently, treating these points as weak regions in the document, we compute its coherence deficiency index (CDI), which is a statistical indicator of the extreme variability in one-dimensional data.

Entailment is a closely related concept to Next Sentence Prediction and has received research stimulation due to revolutionary advances in natural language inference (NLI) models [22]–[24]. These models were trained to detect entailment as classification task, and have been used to infer entailment as an indication of coherence in diverse systems [19], [25], [26]. Entailment is a stricter concept, where the goal is to predict whether a sentence pair (premise and hypothesis) is contradictory, in agreement, or neutral. On the other hand, next sentence prediction task has a comparatively lenient objective, which is to gauge the extent to which the second sentence logically follows the first sentence. We are interested in examining if the information flow through sentence-to-sentence transition is smooth and without any break in continuity. Therefore, BERT NSP score is an intuitive choice for gauging continuity in writing within inter-sentential level.

The merit of our proposed work lies in the innovative way of using BERT NSP signals as indicator to quantify textual coherence and identify regions with weak coherence. Simple and inexpensive, the computation allows us to reuse pre-trained language models and spares us the need to train our own model. Moreover, as BERT NSP task is trained on general English language corpus, the model is readily adaptable to any domain. To the best of our knowledge, this is a first attempt to capitalize on the ability of BERT NSP task to quantify global coherence of the text in zero shot, and identify regions with potential for improvement.

The paper is organized as follows. In Section II, we present the background and related works. In Section III, we describe the methodology used in our research. Sections IV and V are devoted to the approach we propose to detect linear coherence at sentence-level and global coherence at document-level, respectively. We present empirical evidences to support our claims within the sections. We analyze this manuscript with the proposed approach and present our observations in V-C2. Finally, we conclude our study and present future directions in section VI.

II. RELATED WORKS

We relate existing research in computational discourse coherence analysis (CDCA) to our work along two dimensions. Firstly, the approaches - since our approach is a stark departure from the current research trend; and secondly, the tasks - since coherence detection modeled as different tasks mandate a specific evaluation method.

A. APPROACHES FOR COMPUTATIONAL DISCOURSE COHERENCE ANALYSIS

Existing works on computational discourse analysis model the problem as capturing and analyzing interactions between various discourse entities in the text and assessing the thematic progression. Earlier works focused on entity-based approach that captures patterns of entity distribution and interaction in a discourse [5], [27], [28]. Recent works shifted towards neural approaches that employ sequence-to-sequence models, and use word and sentence embeddings to enrich the training set [12], [14], [15], [29].

Entity-based approaches captures coherence on the basis of how discourse entities are distributed across the sentences in the text, and asserts that repetitive mention of the same entity is an indication of a locally coherent text. Pioneering work by Barzilay and Lapata [5] to detect coherence by transforming the text into an entity-grid model, exploiting Centering Theory [3], [30], set an early trend for computational discourse coherence analysis. Entity-grid model is inexpensive and is based on shallow linguistic properties. Several improvements and extensions on this model have been proposed [6], [31], [32]. Later, Guinaudeau and Strube extend the entity-grid model to an entity-graph model, and argue for the proposed model’s computational efficiency and comparable performance [32]. In this model, the entity-graph is first created as a bipartite graph capturing occurrence of entities in sentences of the text, which is subsequently converted to a projection graph of sentences as nodes. Nodes are connected if they share common entities. Local coherence of the text is computed as average outdegree of the projection graph.

Developments in computational discourse analysis methods during the last three years exhibit a revolutionary shift towards deep neural network based architectures. Recurrent Neural Networks, recursive neural networks, sequence-to-sequence architectures, graph recurrent networks, etc. have been explored in recent works [9], [11], [14], [15], [29], [33]–[35]. External resources, such as word embeddings are used to capture contextual and semantic relations between words in text.

Li and Hovy [36] employ distributed sentence representation and implement two approaches using recursive and recurrent neural networks. They evaluate the models for sentence ordering and readability assessment tasks and report improved performance as compared to entity-based approaches. The advance triggered research interest in neural network based approaches [9]–[11] that grew exponentially with evolving neural architectures [12], [14], [15], [29].
Performance of neural network based approaches for computational discourse analysis surpasses that of non-neural, entity-based and lexical models by a wide margin [10]. However, on the flip side, neural network based models are resource intensive and suffer from lack of explainability. These models are trained on sufficiently large datasets, which requires extensive efforts to curate such quality datasets.

B. CDCA TASKS FOR PERFORMANCE EVALUATION

Performance of CDCA models is evaluated on a number of tasks - including Sentence Ordering [5], Sentence Insertion [10], n-way Classification [10], [12], [19], and Coherence Score Prediction [10], [12], [19]. The most common evaluation task is sentence ordering (shuffle test), where the goal is to identify the most coherent order from the set of jumbled sentences. Sentence ordering is designed either as a generative task, where the model aims to find an optimal ordering of sentences that maximizes the coherence [9], [15]–[17], [34], or as a discriminative task, where the model learns to correctly rank the original document higher than some random permutations of the same sentences [5], [10], [18], [19], [32]. Recently, Farag et al. [29] extended the sentence ordering task to sentence pair discrimination task, where the goal is to identify the original sentence pair among its linguistically perturbed variations.

Lai and Tetreault [10] commented that the coherence models trained and tested on the sentence ordering task often over-estimates the success with near-perfect performance, where in real-life such performance is seldom achievable. Laban et al. [37] affirms this view and argue that the sentence ordering task (shuffle test) should be approached in a zero-shot setting, without training the models on the specific task.

We share the serendipity of BERT NSP task for quantitative assessment of discourse coherence of a document with meager computational expense. We evaluate our method in a zero-shot setting, as recommended in [37].

III. COHERENCE DETECTION USING BERT NSP

Van Dijk [2] stipulated that discourse coherence can be linear, i.e., coherence between sequence of sentences, or global, i.e., coherence of the document as a whole. Accordingly, we propose a method to assess both - linear coherence between consecutive sentences and global coherence of the document.

Linear (or Sequential) coherence between consecutive sentence pairs in texts is the lowest level assessment of discourse coherence. We capture the extent of linear coherence in a pair of consecutive sentences using the BERT Next Sentence Prediction (NSP) task. Given a pair of sentences, the BERT NSP task is designed to predict the probability of the second sentence being a valid ‘next sentence’ to the first sentence [22]. BERT NSP task is modeled as a binary classification task that yields two values to represent unnormalized scores (logit) from the model corresponding to the two classes - ‘Next Sentence’ and ‘Random Sentence’. We use the first value, which indicates the score for ‘Next Sentence’.

Treating BERT NSP scores as signal of linear coherence in consecutive sentence pairs, we use elementary signal processing to propose a fast and frugal coherence detection (FFCD) method. We first localize points of unduly weak coherence. Subsequently, treating these points as noisy regions in the document, we compute coherence deficiency index (CDI) to assess global coherence of the document using heuristically defined functions.

We design experiments to establish the effectiveness of BERT NSP and CDI metrics for sentence-level and document-level coherence with an objective to assess the thematic progression of discourse and lack of coherence in the text.

IV. LINEAR COHERENCE IN CONSECUTIVE SENTENCES

We compute BERT NSP score for a sentence pair using the pre-trained ‘bert-base-uncased’ model with a Next Sentence Prediction head. We use the unnormalized score, which gives us an idea about how well the second sentence in a pair of consecutive sentences follows from the first sentence. Shi and Demberg [38] and Kishimoto et al. [39] establish the efficacy of BERT NSP task for discourse relation classification. Based on their analysis, it is apparent that the model is easily transferable to other domains with minimal re-training and fine-tuning. The following example snippet borrowed from Laban et al. [37] demonstrates the efficacy of BERT NSP in capturing linear coherence in consecutive sentence pairs.

Example 1: Figures 1a, 1b, and 1c respectively show texts with coherent sentences, all shuffled sentences, and block shuffled sentences. We compute BERT NSP scores for the consecutive sentence pairs in the three example texts, and plot the scores in Figures 1d, 1e, and 1f.

Consistent with the stipulation by Laban et al. [37], the range of NSP scores in y-axis of the line plots varies based on the extent of incoherence in the three examples. In the text shown in Figure 1b, where all sentences are randomly shuffled, the second sentence pair (sentences 2 and 3) has negative NSP score (Figure 1e), which correctly implies that the extent of coherence at this transition is extremely low. In the text shown in Figure 1c with block shuffled sentences, BERT NSP successfully identifies the point of shuffle, where the first three sentences are pairwise coherent and so are the last three. However, the transition from sentence 3 to sentence 4 lacks coherence (Figure 1f).

Below we empirically demonstrate the effectiveness of BERT NSP task in detecting linear coherence in written discourse.

A. EFFECTIVENESS OF BERT NSP SCORES FOR DETECTING LINEAR COHERENCE

Coherence detection has been modeled as a sentence pair discrimination task, designed to recognize synthetically introduced syntactic and semantic perturbations in the text. Recently, Farag et al. [29] introduced a syntactic perturbation task to assess the robustness of existing models in detecting synthetically induced incoherence.
S. Duari, V. Bhatnagar: FFCD: Fast-and-Frugal Coherence Detection Method

**FIGURE 1.** BERT next sentence prediction scores identify coherent deficient sentence pairs.

Following Farag et al. [29], we model coherence detection in sentence pairs as a pairwise ranking task, where we check the effectiveness of BERT NSP score for accurately detecting various perturbations. We compute pairwise ranking accuracy (PRA), which is the ratio of correct pairwise rankings between a coherent example and its incoherent counterpart. Specifically, we aim to investigate the potency of BERT NSP score for capturing aspects of text implicated in discourse organization.

1) **DATASET FOR INVESTIGATING LINEAR COHERENCE**

We use the Cloze Coherence (CC) data set and Controlled Linguistic Alterations (CLA) data set introduced by Farag et al. [29] for this experiment.

**a: CLOZE COHERENCE (CC) DATASET**

This dataset consists of coherent sentence pairs and corresponding incoherent pairs created by applying syntactic or semantic perturbations. The perturbations are introduced by swapping the sentence order and replacing the second sentence with a random sentence. Therefore each coherent example has two incoherent examples. Based on the transformation, the CC dataset is divided into two sets - cloze_swap and cloze_rand. Each subset consists of 1871 pairs of coherent and incoherent examples.

**b: CONTROLLED LINGUISTIC ALTERATIONS (CLA) DATASET**

This data set is used to evaluate coherence models for detecting controlled set of linguistic changes. Six perturbations are performed for each coherent sentence pair to produce as many examples of incoherent sentence pairs. We perform comparative evaluation on three types of perturbations - random, lexical perturbation, and corrupt pronoun.

Random involves two changes - replacing the first sentence with a random sentence or replacing the second sentence with a random sentence. Lexical perturbation keeps the sentence order intact, and replaces all occurrences of a selected word with a replacement word (e.g., ‘word’ with ‘pen’). Examples of corrupt pronoun are created by keeping the sentence order intact and replacing the subject pronoun in second sentence with another pronoun (e.g., ‘it’ with ‘he’). The CLA dataset contains 30 examples of lexical perturbation and corrupt pronoun, and 60 examples for Random perturbation with 30 examples each for two types of random insertions.

2) **EXPERIMENTAL RESULTS**

We demonstrate the efficacy of BERT NSP score for discriminating coherent examples from incoherent counterparts in Tables 1 and 2. We compare the performance with top-three models as reported in [29]. It is noteworthy that we employ a pre-trained model for BERT NSP score computation whereas the three competing models are based on deep learning and trained for coherence detection.

As shown in Table 1, BERT NSP performance on the cloze_swap set is comparable to $E_{grid,cnn}$ model, and significantly better than the $MTI_{bert}$ and $LCD_{bert}$ models. For the cloze_rand dataset, BERT NSP outperforms the three competing models with a significant margin. For the CLA dataset in Table 2, BERT NSP outperforms the three competing models for the three perturbation types - random, lexical perturbations, and corrupt pronoun. This experiment affirms that BERT NSP score is effective in detecting random insertions.

We do not evaluate our method on the perturbations for swapped sentence pairs as we observed that swapping of
sentences in a pair in this dataset does not always lead to a decrease in coherence. The following example from the CLA data set demonstrates the veracity of our argument.

**Example 2 (Original Coherent Sequence):** Labour is pushing new boundaries in its challenge to the big corporations across the economy. It says the pharma companies are vested interests making excess profits on public health.

**BERT NSP score** = 5.604059332

**SWAPPED SEQUENCE:** It says the pharma companies are vested interests making excess profits on public health. Labour is pushing new boundaries in its challenge to the big corporations across the economy.

**BERT NSP score** = 6.062902451

In the above example, the swapped pair has a higher NSP score than the original pair. A human reader will be able to comprehend both sentence pairs equally well without detecting any incoherence, despite the ambiguous co-reference in the first sentence of the swapped sequence.

We investigated the CLA swap dataset and found 13/30 sentence pairs exhibiting similar pathology. A neural model trained on such dataset would be capable of identifying such cases. Since we favor an unsupervised approach, we are apprehensive about the suitability of the CLA swap data set for the task and thus do not venture to experiment with this data set.

### V. COHERENCE AT DOCUMENT LEVEL

A sequence of linearly coherent sentences does not guarantee a globally coherent discourse, as it is possible to drift from the original theme as the discourse progresses. Therefore, assessing the global coherence at the document level is necessary to gauge the overall discourse quality.

Let \( D = \{s_1, s_2, \ldots, s_n\} \) be a document with \( n \) sentences. Further, let \( b_i \) be the BERT NSP score for the pair of consecutive sentences \( (s_i, s_{i+1}) \). We obtain vector \( B = (b_1, b_2, b_{n-1}) \) of the sentence pair coherence scores corresponding to the document \( D \). Vector \( B \) thus encodes raw signals of coherence between consecutive sentence pairs of \( D \), with higher scores indicating a higher degree of coherence.

Though a coherent text exhibits high sequential coherence for all sentence pairs compared to one with some coherence deficit sentence pairs, it is unrealistic to expect all scores to be identical or uniformly distributed. The thought being communicated by the author, pragmatics of the language, writing style, etc., are some of the factors that induce variation in the scores even in highly coherent and well-written text. Even though consistent high value for all sentence pairs demonstrates coherent thematic progression, the existence of **spikes** and **dips** is inevitable. However, if a dip is sharper than the other dips in the neighborhood, it exposes semantic relation detrimental to coherence between the corresponding sentence pair. Identifying tell-tale dips in the midst of a continuous train of variable signals is the bonafide action for assisting the author in improving text coherence.

The proposed FFCD approach takes cognizance of the aberrant dips, which are noticeable signs of discontinuity in the discourse. Such dips connote events of interest for the purpose of coherence analysis. We propose to detect these events using classical non-parametric outlier detection method based on inter-quartile range [40]. We generate five-number summary of the sentence-permuted (shuffled) text, i.e. sentence pairs 1, 5, 6, and 9, with significantly lower NSP scores showing discontinuity in the discourse quality. We observe that i) there is one significant dip (outlier) in the signal for the original text located at the sentence pair 2 (Figure 2c), highlighted in Figure 2a. ii) there are four outliers in the sentence-permuted (shuffled) text, i.e. sentence pairs 1, 5, 6, and 9, with significantly lower amplitudes compared to the rest. These sentence pairs are also highlighted in the corresponding text.

Spotting the incoherent sentence pairs in the text assists the author to focus attention to regions of weak semantic connectivity.

### A. COHERENCE DEFICIENCY INDEX

Given the sequence of BERT NSP scores for the consecutive pairs of sentences in a document, we study the distribution of outliers to assess the extent of global coherence.

\[ Q_1 \] The five-number summary is a set of descriptive statistics that consists of the five most important sample percentiles: the sample minimum (\( \min \)), the lower quartile or first quartile (\( Q_1 \)), the sample median (\( Q_2 \)), the higher quartile or third quartile (\( Q_3 \)), and the sample maximum (\( \max \)). Inter-quartile range (IQR) is defined as \( Q_3 - Q_1 \).
in the text document. Since the length of the signal is fall-out of the length of the document, we confer distinct treatments for short and medium/long documents. Short texts (signal lengths ≤ 30) warrant treatment appropriate for statistically small samples, while medium and long texts (signal lengths > 30) need to be dealt with techniques pertinent for statistically large samples.

We devise coherence deficiency index (CDI) for both short and medium/long texts, which is a statistical indicator of the variability in one-dimensional data. The proposed index is based on efficient and prudent processing of BERT NSP signals.

**B. COHERENCE DETECTION IN SHORT TEXTS**

Short documents such as news articles, short reports, summary, etc. usually focus on a single topic and there is seldom any segmentation into sections or subsections. These documents yield short-length signal, which are amenable to a quick assessment of global coherence. The number of sentence pairs with low NSP scores is a noticeable indicator of the degree of global coherence in short text.

We conjecture that coherent and well-written text has lower variability in signals and fewer outliers. Accordingly, the ratio of number of outliers to the number of sentence pairs is a competent measure of the extent of text coherence. We define coherence deficiency index for short text (s-CDI), \( \sigma \), for assessing coherence in short text as follows.

\[
\sigma = \frac{k}{n},
\]

where \( k \) is the number of outliers and \( n \) is the number of sentence pairs. The metric \( \sigma \) lies between (0,1] and well-written texts are expected to have a value closer to 0. Farther away the value is from 0, lesser is the extent of coherence in text. In contrast, higher number of outliers admits multiple regions of weak coherence in the text. Further, two or more consecutive outliers are manifestation of a block of text with disputable coherence.

1) EFFECTIVENESS OF S-CDI

We experiment with order discrimination task (shuffle test) based on sentence order permutation as proposed by Barzilay and Lapata [5] to evaluate the performance of the s-CDI metric \( \sigma \). The task is modeled as a pair-wise learning to rank task where the goal is to correctly identify the original text given a pair of original and its sentence-permuted variation. Specifically, we answer the research question: How well does \( \sigma \) discriminate between coherent and incoherent documents?

We approach this problem in a zero-shot setting, as suggested in [37] and apply the proposed \( \sigma \) metric to quantify the coherence in the text. The number of times the original document was ranked higher in coherence than its shuffled versions indicates the efficacy of the metric.

### a: DATA FOR COHERENCE DETECTION IN SHORT TEXT

In this experiment, we use two well-explored datasets - Accidents and Earthquake [5]. Both datasets have been extensively used in previous studies for local and sequential coherence analysis [5], [16], [32], [33], [45].
Table 3: Dataset details. |D|: Number of documents in the collection with pre-defined train and test splits, Savg: average number of sentences per document.

| Dataset   | |D|  | Savg |
|-----------|-----------------|-----------------|
| Accidents | Train | 100 | 100  |
| Earthquakes | Test | 100 | 99   |

Table 4: Result on test set of the earthquakes and accidents datasets.

| Models          | Accidents | Earthquakes |
|-----------------|-----------|-------------|
| Entity Grid [5] | 90.4      | 87.2        |
| Entity Graph [32] | 84.6 | 63.5        |
| Li et al. [16]  | 93.0      | 99.2        |
| ATToOrderNet [33] | 96.2 | 99.8        |
| RankTxNet [45]  | 96.8      | 98.6        |
| σ               | 90.5      | 87.1        |

The Accident dataset consists of 200 documents detailing accident reports, which are written by government officials [5]. The documents in this dataset are short texts, containing an average of 11 sentences each. The Earthquake dataset consists of 199 news articles on natural disasters [5]. Similar to the Accident dataset, the documents in this dataset are also written by experts, thus ensuring coherence and cohesion. The documents in these datasets are short texts, containing an average of 10 sentences each.

Each document in the two data sets is accompanied by 20 sentence-permuted versions, unless the document contains fewer sentences and 20 permutations of sentence ordering is not possible. In such cases, number of sentence-permuted versions is kept as the number of possible distinct permutations. The original documents are considered as coherent examples and the sentence-permuted ones are treated as incoherent examples. An ideal metric should score the original document with higher coherence score than its sentence-permuted variations.

Table 3 summarizes the statistics about the datasets. We use the same Train and Test split for both datasets as introduced in the original paper [5].

b: EXPERIMENTAL RESULTS

We compare the proposed metric σ against five competing methods - Entity Grid [5], Entity Graph [32], approach by Li and Jurafsky [16], ATToOrderNet [33], and RankTxNet [45]. Entity Grid approach and its extension Entity Graph are entity-based approaches, which analyze the entity transition patterns in the discourse. Remaining three approaches, i.e., the approach by Li et al., ATToOrderNet, and RankTxNet, are neural approaches that employ sophisticated deep neural architectures to uncover latent patterns.

Table 4 presents our empirical results against the results reported by the five competing methods. The performance for the five competing methods has been quoted from the respective references. It is evident from Table 4 that σ is unable to outperform state-of-the-art sophisticated deep neural net based models on similar tasks, which is understandable. However, with only a pre-trained model that is not fine-tuned for the task in hand, we were able to achieve comparable performance to that of the entity-grid model and a significantly better results to that of the entity graph model. The proposed approach is fast and frugal as it requires minimal processing and resources, and establishes the efficacy of σ score for capturing local coherence.

C. COHERENCE DETECTION IN LONG TEXTS

Non-short texts (long and medium-length texts) like scholarly articles and reports often span more than 30 sentences. Such texts are usually segmented in semantic units as sections and subsections etc. There is a global theme that drives the text, however the section and subsection focus on inter-related sub-themes. Though a human reader can link together latent sub-themes in structurally well-organized text even if they span over different semantic units, computational methods are not competent to detect such long range dependencies. Therefore, it is impertinent to look for a consistent NSP score throughout the document. Any computational method will detect low coherence scores at paragraph, section, or subsection boundaries. Ergo, non-short texts warrant a distinct treatment for quantitative analysis of discourse coherence.

Raw BERT NSP signal (B) for non-short texts may have weak signal-to-noise ratio due the several reasons including the author’s writing style, structure of the document, segmentation of semantic units like sections, subsections, paragraph, etc. These factors cast a distinctive influence over the variability in raw BERT NSP signal. Therefore, it is imperative to smooth the signal prior to subjecting it to further treatment. We smooth the signal in B by using Kolmogorov-Zurbenko (KZ) filter [46], which is a low pass-filter involving a series of iterations of the moving average (MA) filter. KZ filter is suitable for signal smoothing in our application because it is efficient, robust and optimal.

KZ filter takes two parameters (number of iterations k, and window size m) as input, which have clear interpretations in our case. We set k = 3 for smoothing, which yields an approximately Gaussian-shaped filter [46]. Our choice of window size (m = 3) is motivated by the conjecture that the flow of information from a sentence gets diluted beyond a span of next four-five sentences [47]. The window size of three ensures that we consider four consecutive sentences at a time to assess the pairwise coherence. For example, an outlier score in a window size m = 3 would affect its two non-outlier neighbors. However, if for the same outlier, we smooth the signal using m = 6, the five non-outlier values will mute the effect of the single outlier value.

We pass the raw signal B through KZ filter and denote the vector representing the smoothed signal as S. Since two consecutive coherence values are unlikely to be identical, the smoothed signal is constantly varying in time with frequent spikes and dips. Location of outliers in S guides the author to inspect the quality of written discourse for possible improvement, as mentioned in Section V. It is intuitive that
incoherence in a set of sentences distributed randomly throughout the document is less severe that incoherence spanning a block of consecutive sentences. Thus, cogent analysis of the signal in $S$ is obligatory to assess global coherence. We design a coherence assessment metric based on variability in the data as described below.

Let $O = \{o_1, o_2, \ldots, o_k\}$ be the sequence of outlier indices detected from the smoothed signal $S$, $k$ being the number of outliers, and $o_i$ the index of the $i^{th}$ outlier. We use $o_0$ and $o_{n+1}$ to mark the beginning and end of the sequence $O$, and set the corresponding values to 0 and $n + 1$. We define inter-outlier interval as the distance between indices of two successive outliers (i.e., $o_{i+1} - o_i$) and compute the following:

$$\hat{O} = \{(o_1 - o_0), \ldots, (o_{m+1} - o_m)\}. \quad (2)$$

The inter-outlier interval betokens the distribution of outliers in the document. A random distribution of interval represents weak coherence arising occasionally in the text. However, consecutive occurrence of outliers generates small intervals and indicates poorly constructed text segment with low semantic connectivity.

Coherence deficiency index $\xi$ for medium and long text (L-CDI) is defined as the coefficient of variation of the distribution of inter-outlier intervals:

$$\xi = \frac{\text{stdev}(\hat{O})}{\text{mean}(\hat{O})}, \quad (3)$$

where $\text{stdev()}$ and $\text{mean()}$ are functions to compute standard deviation and mean, respectively.

It is noteworthy that unlike $\sigma$, L-CDI metric $\xi$ for non-short text lies in a different range of values ($0 \leq \xi < \infty$). Coherent texts will have $\xi$ closer to the ideal value 0, as they are expected to have no (or fewer) outliers. On the other hand, texts with weaker coherence will have higher values of $\xi$ as they throw more outliers. Interestingly, the computation penalizes signals with consecutive outliers, which signifies a block of weakly connected text.

1) EFFECTIVENESS OF $\xi$ IN CAPTURING GLOBAL COHERENCE

We design an experiment to assess the effectiveness of $\xi$ metric in capturing global coherence of long documents. Specifically, we aim to answer the question: Whether $\xi$ sufficiently discriminate discourse quality?

a: DATA

We use the ‘ACL2017’ dataset from PeerRead collection [48]. The dataset consists of manuscripts submitted to 2017 edition of the ACL conference, where the author and at least one reviewer has agreed to include the manuscript and the review in this dataset. Each document in the dataset is therefore accompanied by a corresponding review file that contains reviewers’ comments and their assigned scores for various assessment criteria, like ‘Clarity’, ‘Recommendation’, ‘Soundness’, etc. on a scale of 1-5 (5 being outstanding, however no documents were assigned a clarity score of 1). As we aim to assess the discourse quality, we focus on the ‘Clarity’ score.

b: DATASET QUALITY

Many documents in this dataset contain only a single review. Specifically, out of 137 total documents, 39 documents have a single review and only 31 documents have multiple reviews with unanimous clarity score. Rest of the 67 documents (≈50%) have mixed clarity score, with the disagreement sometime ranging between 2 to 5. This observation confirms the subjective nature of discourse comprehension by reader.

We analyze documents which were assigned a single or unanimous clarity score of 2 and 5 by the reviewers. We are left with a dataset of only 24 documents where 19 documents have clarity-5 and only 5 documents have clarity-2. We also add documents with a minimum clarity score of 2, which increases the clarity-2 corpus to 18 documents.

c: EXPERIMENTAL RESULTS

We perform basic cleaning of the text. We remove the abstract, references, acknowledgements, appendices, examples, equations, and tables and figures along with their captions. We also remove sentences with too few (<5) words and too many special characters and mathematical symbols to avoid sentences with inline equations and mathematical notations. We retain the remaining sentences for further analysis.

We compute the BERT NSP score for sentence pairs in the documents. We however do not include the consecutive sentence pairs at section and subsection boundaries, as coherence at such transitions is expected to be lower leading to detection of outliers which are false positive. We compute $\xi$ score for each document and assess the discrimination capability of the score by plotting $\xi$ scores for the two classes of documents in Figure 3.

We observe that the $\xi$ score for documents with clarity score 5 is lower than that of documents with clarity score 2. There is a minor overlap, where two (2) documents with clarity score 2 (doc ids 67 and 318) have a smaller $\xi$ score. We investigated and found out that these documents lack reviewers’ agreement on the assigned clarity score and at least one of the reviewers assigned a high clarity score of 5 to these documents. We also have three (3) documents

![FIGURE 3. Coherence score for clarity 2 and clarity 5 documents computed using $\xi$. The dotted line distinguishes the two classes of documents, where $\xi \approx 0.55.$]
with clarity-5 (doc ids 87, 557, 805) which have a higher $\xi$ score, which is higher than many clarity-2 documents. However, the number of documents with such behavior is small, i.e., 5/37 documents.

The above observation establishes the efficacy of the fast and inexpensive $\xi$ score in distinguishing the discourse quality. Though the performance of coherence deficiency index is not perfect and the inference is deduced from a small sample of data, we endorse the competence of the proposed approach in facilitating a quick assessment of the discourse quality.

2) COHERENCE ANALYSIS OF THIS MANUSCRIPT

We pre-process this manuscript as explained in Section V-C1.c, and assess its quality using $\xi$ score. This helped us to quickly gauge and improve the discourse by analysing the outlier signals. $\xi$ score of the submitted version is $\approx 0.47$ after refinement. We observed that few outliers are false positives, where the machine failed to recognize latent coherence in the sentence pair. Following is an example of such false positive. The sentence pair in the example is from the last paragraph of Section VI of this manuscript.

Example 4: The serendipity of the proposed FFCD method lies in its ability to fairly reflect the coherence in sentence pairs, without retraining and fine-tuning. The fast-nass and frugal-ity of the computation of CDI makes it suitable for online application.

NSP Score = $5.655398$

If the two acronyms are replaced by the full forms, the NSP Score increases to 6.408267 and the pair is no more flagged as coherence-deficient.

Outliers in the signal allude to the likelihood of weak semantic associations, but the author of the text is the ultimate judge and decides the rectitude of the machine. Since pre-trained language models do not understand language in the sense that humans do, the outlying amplitudes of the BERT NSP signals are indicative and so is the CDI. We fully subscribe to the human-in-the-loop view of coherence detection mechanisms, since even in the human world coherence evaluation of discourse is a subjective task.

VI. CONCLUSION AND FUTURE DIRECTION

We study the efficacy of BERT NSP for unraveling coherence of written discourse, and propose a fast-and-frugal coherence detection (FFCD) method based on BERT NSP.

Our empirical study establishes that BERT NSP score is capable of assessing sequential coherence in consecutive sentence pairs in a text. We subject the scores to simple signal processing and statistical techniques to localize regions of weak coherence for short and long texts, and compute coherence deficiency index (CDI). The proposed FFCD method underperforms for sentence ordering task as compared to the SOTA; however, it can be used to quickly gauge the discourse quality and locate regions of weak coherence.

The serendipity of the proposed FFCD method lies in its ability to fairly reflect the coherence in sentence pairs without retraining and fine-tuning. The fast-ness and frugal-ity of the computation of CDI make it suitable for online application. We envisage that this work sets the direction for further scrutiny of pretrained language models trained for next sentence prediction or entailment task for their potential in assessing discourse coherence. More refined analysis of the variability in the BERT NSP signal (e.g. Fano factor, auto-regression, spectral analysis etc.) can potentially yield superior methods with better performance.
[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.

[23] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “RoBERTa: A robustly optimized BERT pretraining approach,” 2019, arXiv:1907.11692.

[24] S. Aroca-Ouellette and F. Rudzicz, “On losses for modern language models,” in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 4970–4981.

[25] N. Dzieli, E. Kamalloo, K. Mathewson, and O. Zaiane, “Evaluating coherence in dialogue systems using entailment,” in *Proc. Conf. North*, 2019, pp. 3806–3812.

[26] S. Storks and J. Chai, “Beyond the tip of the iceberg: Assessing coherence of text classifiers,” 2021, arXiv:2109.04922.

[27] M. Lapata and R. Barzilay, “Automatic evaluation of text coherence: Models and representations,” in *Proc. IJCAI*, vol. 5, 2005, pp. 1085–1090.

[28] R. Barzilay and M. Lapata, “Modeling local coherence: An entity-based approach,” *Comput. Linguistics*, vol. 34, no. 1, pp. 1–34, 2008.

[29] Y. Farag, J. Valvoda, H. Yannakoudakis, and T. Briscoe, “Analyzing neural discourse coherence models,” in *Proc. 1st Workshop Comput. Approaches Discourse*, 2020, pp. 102–112.

[30] B. J. Grossz, S. Weinstein, and A. K. Joshi, “Centering: A framework for modeling the local coherence of discourse,” *Comput. Linguistics*, vol. 21, no. 2, pp. 203–225, 1995.

[31] M. Elson and E. Charniak, “Extending the entity grid with entity-specific features,” in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, pp. 125–129, 2011.

[32] C. Guinaudeau and M. Strube, “Graph-based local coherence modeling,” in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics, Long Papers*, vol. 1, 2013, pp. 93–103.

[33] B. Cui, Y. Li, M. Chen, and Z. Zhang, “Deep attentive sentence ordering network,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 4340–4349.

[34] L. Logeswaran, H. Lee, and D. Radev, “Sentence ordering and coherence modeling using recurrent neural networks,” in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 5285–5292.

[35] Y. Yin, L. Song, J. Su, J. Zeng, C. Zhou, and J. Luo, “Graph-based neural sentence ordering,” in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 5387–5393.

[36] J. Li and E. Hovy, “A model of coherence based on distributed sentence representation,” in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 2039–2048.

[37] P. Laban, L. Dai, L. Bandarkar, and M. A. Hearst, “Can transformer models measure coherence in text? Re-thinking the shuffle test,” 2021, arXiv:2107.03448.

[38] W. Shi and V. Demberg, “Next sentence prediction helps implicit discourse relation classification within and across domains,” in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 5790–5796.

[39] Y.kishimoto, Y. Murawaki, and S. Kurohashi, “Adapting BERT to implicit discourse relation classification with a focus on discourse connectives,” in *Proc. 12th Lang. Resour. Eval. Conf.*, Marseille, France: European Language Resources Association, May 2020, pp. 1152–1158.

[40] J. W. Tukey, *Exploratory Data Analysis*, vol. 2, Reading, MA, USA: Addison-Wesley, 1977.

[41] F. T. Liu, K. M. Ting, and Z.-H. Zhou, “Isolation forest,” in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 413–422.

[42] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining*, Portland, OR, USA, Aug. 1996, vol. 96. no. 34, pp. 226–231.

[43] D. V. Hinkley, “Inference about the change-point in a sequence of random variables,” *Biometrika*, vol. 57, no. 1, pp. 1–17, Apr. 1970.

[44] C. Mikkö and H. Lajos, *Limit Theorems in Change-point Analysis*. Hoboken, NJ, USA: Wiley and Sons, 1997.

[45] P. Kumar, D. Brahma, H. Karnick, and P. Rai, “Deep attentive ranking networks for learning to order sentences,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 8115–8122.

[46] I. G. Zubkenko, *The Spectral Analysis of Time Series*, vol. 2, Amsterdam, The Netherlands: North Holland, 1986.

[47] M. Abdolahi and M. Zahedi, “A new model for text coherence evaluation using statistical characteristics,” *J. Electro. Comput. Eng. Innov.*, vol. 6, no. 1, pp. 15–24, 2017.

[48] D. Kang, W. Ammar, B. Dalvi, M. Zuylen, S. Kohlmeier, E. Hovy, and R. Schwartz, “A dataset of peer reviews (PeerRead): Collection, insights and NLP applications,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, (Long Papers), vol. 1, 2018, pp. 1647–1661.