Journal of the Brazilian Computer Society, 2025, 31:1, doi: 10.5753/jbcs.2025.3705

Semantic Coherence of Short Text at the Word Level

Osmar de Oliveira Braz Junior [**State University of Santa Catarina** | ***osmar.braz@udesc.br***]
Renato Fileto [**Federal University of Santa Catarina** | ***r.fileto@ufsc.br***]

State University of Santa Catarina (UDESC), Av. Madre Benvenuta, 2007, Florianópolis, 88.035-901, SC, Brazil

Received: 26 September 2024 • **Accepted:** 26 May 2025 • **Published:** 25 June 2025

Abstract Most text coherence models proposed in the literature focus on sentence ordering and semantic similarity of neighboring sentences. Thus, they cannot be applied to documents with just one sentence and do not properly look at incoherences caused by particular words. This work, on the other hand, focuses on word coherence in short texts. It proposes a framework called COHEWL (COHErence at Word Level) for assessing short document coherence at the word semantic level. COHEWL also support contrastive data generation by exchanging particular words with other ones that may fit in the context of the respective documents. Experiments with single-sentence questions typical of QA in Brazilian Portuguese and English were conducted. BERT, properly trained for a new task proposed in this paper – discriminating original documents from those with a changed word – achieves accuracy between 80% and 99.88%. However, our experimental results did not show relevant correlations of the BERT Masked Language Model (MLM) word prediction rank with coherence (or incoherence) measures calculated as average similarities (or distances) between BERT embeddings of text changed by the predicted words. In addition, in our manually created corpus of coherent and incoherent questions about data structures, coherence measures based on a topic model built from a few documents of the same domain discriminate coherent documents from incoherent ones with much higher precision than the coherence measures derived from BERT embeddings.

Keywords: Semantic Coherence, Short Text, Word Semantics, Language Models, Contextualized Embeddings

1 Introduction

Coherence is essential for discourse understanding. It distinguishes easy to understand text or speech from confusing and/or inconsistent ones [Halliday and Hasan, 2014]. A coherent text or speech has semantically compatible components (words, sentences, etc.), allowing congruent interpretation. Incoherences, on the other hand, can cause interpretation difficulties, compromising the communication performance of humans, conversational agents, intelligent tutors, and Question Answering (QA) systems.

Differently from this work, whose focus is on the semantic coherence of words in short texts, current models for measuring text document coherence are usually based on the similarity of adjacent sentences [Barzilay and Lapata, 2008]. In addition, the vast majority of the coherence classification models from the literature [Vakulenko *et al*., 2018; Joty *et al*., 2018; Mesgar and Strube, 2018; Xu *et al*., 2019; Bao *et al*., 2019; Farag and Yannakoudakis, 2019; Moon *et al*., 2019] are evaluated using sentence order discrimination (also known as Shuffle Test) [Barzilay and Lapata, 2008]. This task aims to distinguish original documents from the ones with permuted sentences[1]. Classification models for the Shuffle Test may learn to recognize documents with permuted sentences instead of adequately assessing coherence [Laban *et al*., 2021]. Therefore, these models only apply to texts with several sentences and do not address semantic coherence at the word level, i.e., semantic adequacy and compatibility of words in short documents.

In this work, we consider only short text documents, consisting of a single sentence, which can be simple (i.e., with one verb) or compound (i.e., with more than one verb). The corpora used in our experiments include only short documents having between 10 and 12 words on average, depending on the dataset. Thus, in this work, we use the terms short text, sentence, and document as synonyms. The following examples illustrate how different words can make such short texts more or less coherent:

1. *"How to **push** elements in a stack?"*
2. *"How to arrange elements in a stack?"*
3. *"How to organize elements in a stack?"*
4. *"How to store elements in a stack?"*
5. *"How to find elements in a stack?"*

Notice that just one word is changed in the short texts exemplified above. Short text 1 uses the verb *push* to precisely designate the operation that inserts an element on the top of a *stack*. In the field of data structures, it is considered more coherent than short texts 2 to 4, which use the less specific verbs *arrange*, *organize*, and *store*. In contrast, the change to the verb *find* makes short text 5 less coherent than the other ones and with another meaning.

Words that are not suitable for specific contexts can compromise coherence and proper interpretation. Thus, automatic detection and solving incoherence at the word level during text pre-processing could benefit downstream tasks that depend on proper interpretation, such as QA. Neverthe-

[1] Many works consider original documents coherent and modified ones incoherent. In fact, this assumption is not necessarily true, as some original documents (e.g., created by humans) can be incoherent, while certain changes can turn incoherent documents into coherent ones, though this rarely happens by chance. Producing supposedly less coherent versions of documents by permuting sentences has been criticized for using criteria that do not match the human perception of coherence [Mohiuddin *et al*., 2021].

less, semantic coherence was not sufficiently investigated yet, particularly the coherence of words in short documents.

This work introduces the COHEWL (*COHE*rence at *W*ord *L*evel) framework for assessing the semantic coherence of words in short documents. It allows generating contrastive data, training, and comparing the performance of alternative models to classify and measure semantic coherence, using state-of-the-art technologies such as Language Models (LM) like BERT [Devlin *et al*., 2019] and topic coherence models [Churchill and Singh, 2022]. Our proposal to generate contrastive data (changed documents) is to randomly choose words to be exchanged by other, which can be proposed by humans or automatically predicted by the BERT Masked Language Model (MLM), among others.

Experiments with short texts containing single-sentence questions typical of QA in Portuguese and English showed that BERT allowed accuracies between 80% and 99.88% when adequately trained for the task of discriminating original documents from the ones with a word changed. We also calculated (in)coherence measures based on averaged similarities (or distances) between BERT embeddings. These embeddings include embeddings of words with embeddings of other words in the same document or embeddings of the whole document (produced by pooling). However, these (in)coherence measures are not necessarily better for original documents than for documents with a word changed. In addition, these coherence measures of changed documents have no relevant correlation with the confidence rankings of the BERT MLM word predictions used to change the documents. On the other hand, on a corpus of questions about data structures that we have built, topic coherence measures are higher for the vast majority of coherent documents than for the respective documents with a word manually changed to make them less coherent.

The major contributions of this paper are:

1. the proposal of a new task that consists of discriminating original documents from those with a word changed;
2. accurate document classifiers produced by fine-tuning BERT using distinct data and configurations of hyperparameters;
3. semantic (in)coherence measures based on average similarities (or distances) of word embeddings with word or sentence embeddings; and
4. comparison of several variations of these (in)coherence measures and topic coherence measures;
5. a dataset of coherent and incoherent single-sentence questions typical of QA in a learning environment in the domain of data structures;
6. the "COHEWL" framework[2] to compare alternative methods for generating contrastive data and assessing short documents semantic coherence at the word level.

This paper is organized as follows. Section 2 provides the foundations needed to understand it. Section 3 discusses related work. The Section 4 presents the COHEWL workflow and its significant tasks, which can be performed using alternative methods. Section 5 reports the experiments realized to evaluate the proposal. Then, Section 6 presents and discusses experimental results of classifiers, Section 7 presents the results involving coherence measures based on BERT embeddings, and Section 8 presents the results related to topic coherence measures. Finally, Section 9 draws the conclusions and future work.

2 Foundations

Discourse is a combination of spoken or written fragments that allows communication [Wang and Guo, 2014]. Text refers to written discourse. Wang and Guo [2014] apud De Beaugrande and Dressler [1981] place coherence among the seven desired characteristics of a discourse. Coherence distinguishes consistent discourses that make sense from those that do not, have contradictions or failures in the composition of ideas [Wang and Guo, 2014].

A semantically coherent discourse consists of groupings or chains of textual components (words, sentences, etc.) where the meanings of its components are compatible, enabling consistent interpretation. Semantically incoherent texts, on the other hand, have components located close to each other in the text but that have incompatible meanings, i.e., are semantically distant from each other or from the context in which they appear (e.g., sentence, paragraph, document as a whole) [Braz Jr. and Fileto, 2021]. Koch and Travaglia [2021] apud Van Dijk *et al*. [1983] define semantic coherence as the meaning compatibility between adjacent elements (local) or among all the elements in a text document (global).

2.1 Semantic Coherence Models

Most models from the literature evaluate local coherence [Foltz *et al*., 1998; Barzilay and Lapata, 2008; Mesgar and Strube, 2018; Muller *et al*., 2019; Moon *et al*., 2019; Das, 2019; Xu *et al*., 2019; Shen *et al*., 2021]. Evaluating global coherence can be too expensive or unreasonable for lengthy documents. According to Wadud and Rakib [2021], the first model to measure coherence was proposed by Foltz *et al*. [1998]. They used Latent Semantic Analysis (LSA) to create a consolidated vectorial representation of sentences. Equation 1 presents their $C_{cos}(D)$ coherence measure for a document D with n sentences, calculated as the average of the cosine similarities (cos) between vectors representing adjacent sentences (s_i, s_{i+1}).

$$C_{cos}(D) = \frac{1}{n-1} \sum_{i=1}^{n-1} cos(\tilde{s}_i, \tilde{s}_{i+1}) \quad (1)$$

In this work (Section 4.2), we adapt Equation 1 in alternative ways to calculate local and global (in)coherence[3] Thus, it varies in the interval $[0, 1]$ (with 1 indicating maximal coherence). For calculating incoherence measures, which vary in the interval $[0, \infty]$ (with ∞ indicating no coherence), we use some distance (e.g., Euclidean, Manhattan) of short documents. However, other models could also be considered. Entity-based coherence models [Barzilay and Lee, 2004;

[2] Source code available at `https://github.com/osmarbraz/cohewl/`

[3] Our coherence measures are averaged similarities (e.g. cosine similarity).

Barzilay and Lapata, 2008] determine if distinct relevant entities throughout the discourse are locally concentrated while maintaining a thread of ideas. The neural coherence model of Mesgar and Strube [2018] captures the semantic flow of adjacent sentences in a text. Coherence relation models [Muller *et al*., 2019; Moon *et al*., 2019; Das, 2019] evaluate syntactic or meaning dependences between sentences. They consider the relation of ideas in a predefined interval around each sentence. Finally, some recent models [Xu *et al*., 2019; Shen *et al*., 2021] calculate the document coherence score as the average of the local coherence scores between consecutive sentences.

Coherence models can be discriminative or generative [Li and Jurafsky, 2017]. Discriminative models learn to distinguish coherent from incoherent discourse, while generative ones produce coherent text. Most coherence models from the literature are discriminative and try to distinguish human-written documents (original) from their automatically modified versions (changed pseudo-documents) [Lin *et al*., 2011]. The most popular task in this field is the Shuffle Test [Laban *et al*., 2021] (also known as sentence order discrimination), introduced by Barzilay and Lapata [2008]. It tries to distinguish original documents from their 20 versions changed by randomly permuting sentences. However, this task does not necessarily handle coherence (just distinguish original documents from changed ones), can not be applied to short text documents with just one sentence and do not consider coherence at the granularity of individual words. In this work we have a different approach: evaluating the semantic coherence of words in short text.

2.2 Topic Coherence Measures

We envisioned that, besides average similarities or distances between embeddings, topic coherence measures can be competitive to measure the coherence of short documents at the word granularity. Thus, another approach investigated in this work builds a topic model for a corpus in a particular universe of discourse or domain and uses it to calculate topic coherence measures. A topic model is an unsupervised mathematical model that takes a set of documents as input and returns a set of topics that accurately and coherently represent the set [Churchill and Singh, 2022, 2023]. Topic coherence is based on the frequency of co-occurrences of words (topics) in a reference corpus and has been shown to correlate with human assessments. For a comprehensive review of topic coherence measures, please refer to Röder *et al*. [2015].

We investigate in this work the use of the topic coherence measures C_{UCI}, C_{NPMI}, C_V and C_{UMass} as alternatives to measure the coherence of short documents. Equation 2 defines C_{UCI} [Church and Hanks, 1989] as the average Pointwise Mutual Information (PMI–Equation 3) of all word pairs. Probabilities are estimated based on word co-occurrence counts in a sliding window that moves over an external reference corpus [Newman *et al*., 2010].

$$C_{\text{UCI}} = \frac{2}{N \cdot (N-1)} \cdot \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \text{PMI}(w_i, w_j) \quad (2)$$

$$\text{PMI}(w_i, w_j) = \log \frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot P(w_j)} \quad (3)$$

Equation 4 defines C_{NPMI} [Aletras and Stevenson, 2013], an improvement over C_{UCI} that uses Normalized PMI (NPMI–Equation 5) and varies in the interval $[-1, 1]$ [Bouma, 2009].

$$C_{\text{NPMI}} = \frac{2}{N \cdot (N-1)} \cdot \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \text{NPMI}(w_i, w_j) \quad (4)$$

$$\text{NPMI}(w_i, w_j) = \frac{\text{PMI}(w_i, w_j)}{-\log P(w_i, w_j)} \quad (5)$$

Equation 6 defines C_V, which employs a variation of NPMI with sliding window of size 110 [Röder *et al*., 2015] to calculate the cosine similarities between a context vector $\overrightarrow{v_c}$ (Equation 7) and each word vector $\overrightarrow{v_i}$ (Equation 8).

$$C_V = \frac{1}{N} \cdot \sum_{i=1}^{N} cos(\overrightarrow{v_i}, \overrightarrow{v_c}) \quad (6)$$

$$\overrightarrow{v_c} = \sum_{i=1}^{N} \overrightarrow{v_i} \quad (7)$$

$$\overrightarrow{v_i} = \{\text{NPMI}(w_i, w_j)^\gamma, \text{NPMI}(w_i, w_j)^\gamma, \ldots, \text{NPMI}(w_i, w_j)^\gamma\} \quad (8)$$

Equation 9 defines C_{UMass}, an asymmetrical confirmation measure between word pairs [Mimno *et al*., 2011].

$$C_{\text{UMass}} = \frac{2}{N \cdot (N-1)} \cdot \sum_{i=2}^{N} \sum_{j=1}^{i-1} \log \left(\frac{P(w_i, w_j) + \epsilon}{P(w_j)} \right) \quad (9)$$

3 Related Work

Figure 1 presents our classification of coherence models from the literature according to traits that we have observed in our literature review. It is intended to help contextualize our research and compare related work. Regarding the coherence type, our focus is in the semantic coherence, with an additional emphasis on a discriminative approach. Most related works analyze semantic coherence at the sentence granularity, i.e., exploiting semantic similarities between sentences. However, some works, like ours, consider the similarities between words, and a few other between sliding windows of text. The coherence scope, on the other hand, refer to the text portions considered: the whole document (global) or just close or adjacent sentences or words (local). Finally, the coherence model base refers to the foundations employed for coherence representation and analysis: relation-based, entity-based or topic-based.

Relation-based models postulate that sentences or groups of words within sentences possess systematic relationships with neighboring sentences [Wang and Guo, 2014]. On

Table 1. Comparison summary of related work on semantic coherence

Work	Analysis Granularity	Analysis Scope	Analysis Task	Base Model	Technologies	Embeddings
Mesgar and Strube [2018]	Sentence	Local	Classification	Relation	LSTM combined with information in embeddings	Static word embeddings Zou *et al*. [2013]
Joty *et al*. [2018]	Word (Relevant entities)	Local	Classification	Entity	Convolutional neural networks over transition representations in entity grid	Glove
Vakulenko *et al*. [2018]	Sentence & word (entity)	Local	Measurement	Relation	Knowledge graph in background	Glove & Word2vec
Bao *et al*. [2019]	Sentence and word	Local & Global	Classification	Topic	Bi-GRU and language model for detection semantic errors	Word2vec
Farag and Yannakoudakis [2019]	Sentence	Global	Classification	Topic	Bi-LSTM based hierarchical model and grammar types	Glove
Moon *et al*. [2019]	Window	Local & Global	Classification	Relation & Topic	Bi-LSTM unifying intentional structure, discourse relations, topics and attention	Word2vec & ELMo
Nie *et al*. [2019]	Sentence	Global	Classification Multiclass	Relation	LM and discourse markers	BERT
Reimers and Gurevych [2019]	Sentence	Global	Measurement	Relation	LM in a Siamese network with addition of a grouping operation	BERT
Wang *et al*. [2019]	Sentence	Local	Classification, Measurement	Relation	Reward function in reinforcement learning and antagonistic training	Glove
Xu *et al*. [2019]	Sentence	Local	Classification	Relation	MLP and characteristics of concatenated embeddings	Glove
Sarzynska-Wawer *et al*. [2021]	Word	Global	Classification	Relation	LSTM and SVM	ELMo
COHEWL	**Word**	**Local & Global**	**Classification, Measurement**	**Relation & Topic**	**LM and Topic Coherence measures**	**BERT**

the other hand, entity-based models track relevant entities throughout the discourse, as originally proposed by Barzilay and Lee [2004] and extended by Barzilay and Lapata [2008]. Topic-based models focus on a single thematic domain or subject, wherein identical or semantically related words tend to appear in the same sentence or proximate sentences, as per the principles posited by Halliday and Hasan [2014].

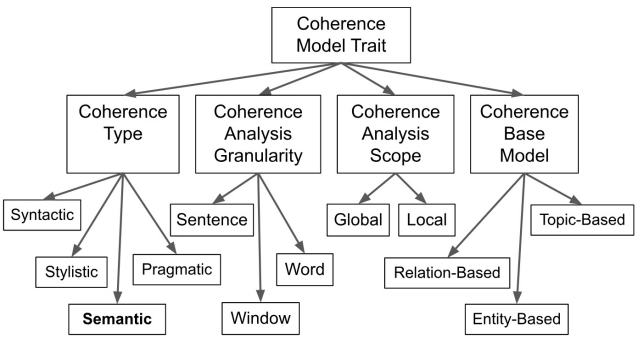

Figure 1. Characteristics of coherence models

Table 1 provides a comparative summary of proposals selected from the recent literature about semantic coherence analysis. The works are ordered chronologically in the table rows. Since all the selected studies focus on semantic coherence, this characteristic has not been included in the table. The second, third and fifth columns of Table 1 refer to the granularity, scope and base model, respectively, as described previously. The fourth column, **"Analysis Task"**, states if the proposed method is for coherence classification and/or measurement. The sixth column, **"Technologies"**, indicates the technologies employed by each proposal. Finally, the kind of **"Embeddings"** are indicated in the last column.

Notice that several works address use features extracted from sentences [Nie *et al*., 2019; Reimers and Gurevych, 2019] or whole documents [Vakulenko *et al*., 2018; Joty *et al*., 2018; Mesgar and Strube, 2018; Wang *et al*., 2019; Xu *et al*., 2019]. In addition, most works address coherence classification and just a few ones coherence measurement.

Regarding the granularity, the vast majority of the proposals use documents with permuted sentences as contrastive data for training or fine-tuning and evaluation with the Shuffle Test. Most works employ static embeddings (i.e., Glove, Word2vec), but there is a recent trend to use contextualized ones (i.e., BERT, ELMo). These embeddings enable identifying relationships between adjacent sentences [Nie *et al*., 2019] or finding pairs of semantically similar ones [Reimers and Gurevych, 2019]. Only Sarzynska-Wawer *et al*. [2021] aim to investigate the impact of words on coherence, determining which words contribute the most to the positive or negative classification of interviews with patients who have schizophrenia and healthy people.

Our work focuses on the coherence of short documents based on the semantic compatibility of their words. To the best of our knowledge, it is the first one to investigate semantic coherence classification and measure using contrastive data produced by exchanging words with other ones that can fit the textual context. In the experiments, we employ the BERT MLM to predict these words and fine-tune BERT for the task of distinguishing original documents from the ones with a word changed. We also measure coherence in different ways, using BERT embeddings of words and sentences (obtained by pooling word embeddings) or topic coherence measures, and then compare the results.

4 The COHEWL Framework

We developed the COHEWL framework to make experiments with alternative methods for changing document words to generate contrastive data, classify and measure the coherence of original and changed documents and then analyze the results.

Figure 2 illustrates the proposed workflow. The process begins by using a NLP tool (e.g., spaCy) to segment N original documents (OD) into their sentences and POS-tag their token. These tasks are necessary to allow subsequent selection of verbs and nouns to change in sentences in the generation of contrastive data for experiments, and for calculating coherence measures based only in verbs and nouns. Then, the task `Generate Contrastive Data` randomly selects a word w in each OD to generate $N_c \geq 1$ changed documents (CD), i.e., modified versions of OD, each one with a distinct word in the place of w. This can be done manually or automatically by, for example, changing w for the token $[MASK]$ to create $maskedOD'$, submitting to the BERT MLM and selecting the N_c predicted words ($pred$) for $[MASK]$ with the highest confidence to use in the place of w in each CD. The pre-processed documents, which include the N original documents (OD) and N_c changed documents (CD) can be used to train `Coherence Classifiers` (we have done it by fine-tuning BERT for the task of discriminating ODs from CDs in our experiments) and for `(In)coherence measurement` (what we have done by averaging embedding similarities (distances) or using a topic coherence model in our experiments). The primary tasks of this process are described in more detail in the following.

4.1 Pre-processing and Contrastive Data Generation

Figure 3 illustrates the generation of contrastive data for the OD *"How to push elements in a stack ?"*. Due to space limitations in this figure we consider $N_c = 10$. In the figure, after OD segmentation and POS-Tagging, the verb ($push$) is randomly substituted by $[MASK]$. Then, the OD with the masked verb is submitted to the BERT MLM to predict N_c words to substitute this verb. In our experiments, a $VERB$ is randomly selected and masked in each OD to be changed by BERT MLM and create $N_c \geq 0$ CD, i.e., versions of each OD with the verb changed with a different word. If no verb is found, the first $NOUN$ is selected, and if neither exists, an auxiliary verb is chosen. Then, the $maskedOD'$ is submitted to BERT MLM, the embeddings of the last layer are recovered, normalized and their probabilities calculated. Then, the N_c top predictions for $[MASK]$ are retrieved to generate the N_c changed versions CD of the original document OD.

Regarding word prediction to substitute the masked one, an important question arises: are the worst models potentially better suited for generating incoherent words, as they might produce more diverse or unexpected replacements? This is particularly relevant when attempting to generate contrastive data that emphasize incoherence. We observed a variety of subtle phenomena in our experiments, such as the prediction of words ($arrange$, $organize$, $store$, put, etc.) with more general meaning than the original ($push$). They make the text less coherent, at least for the domain of data structures. Other words predicted (e.g., $represent$, $sort$) change the meaning of the text, despite being coherent in the context. Thus, we think that more discussion and research are needed to assess how different levels of model quality affect the generation of incoherent candidates, especially in contrastive learning contexts. We leave this to future work due to limited space and to maintain the focus of this paper.

Equation 10 defines the extension of a document set $DocSet$ with contrastive data, i.e., $N_c \geq 1$ CDs for each $OD \in DocSet$, by changing one word in the OD with a distinct word predicted to its place by the BERT MLM.

$$C(DocSet) = DocSet \cup Changes(DocSet, N_c) \quad (10)$$

4.2 Coherence Evaluation

The original documents (OD) and the derived contrastive data (CD) are used to fine-tune classifier of original documents versus documents that had a word changed. Our experiments used BERT and considered the classification token ($[CLS]$) of its last layer as the document class discriminator [Devlin *et al*., 2019]. The generation of embeddings to calculate (in)coherence measures uses the pre-trained BERT in the evaluation mode. Following recommendations from Devlin *et al*. [2019], we do not use all BERT layers, but only the concatenation of the last 4.

In this work, we propose distinct functions derived from Equation 1 [Foltz *et al*., 1998] to calculate semantic (in)coherence measures for each document in $C(DocSet)$, namely: CAW_m, CWP_m and CG_m. These functions can

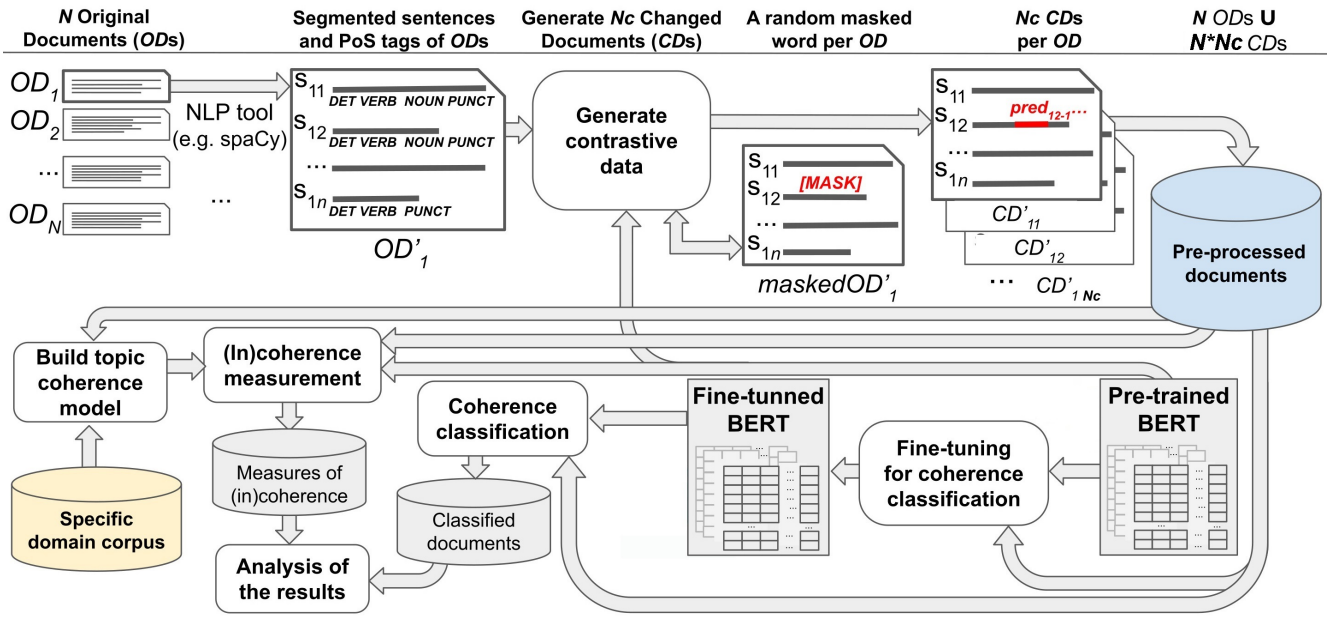

Figure 2. The COHEWL major modules and workflow

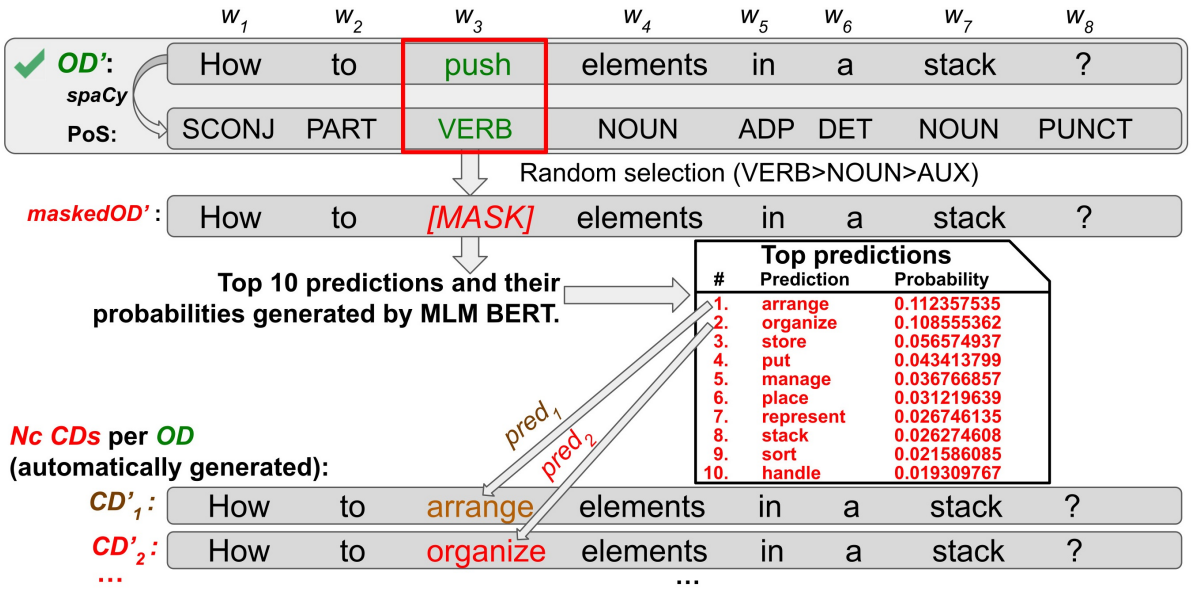

Figure 3. Example of contrastive data (ODs) generation for $N_c = 10$

use any measure m of similarity or distance between embeddings to calculate coherence or incoherence, respectively. In the equations, D represents a document and k represents the number of words in D. Equation 11 presents the function $CAW_m(D)$ (Coherence of Adjacent Words) which corresponds to the average of the similarities or distances (m) between adjacent word (w_i, w_{i+1}) of D. Equation 12 presents $CWP_m(D)$ (Coherence of all Word Pairs), the average similarities/distances of all word pairs w_i e w_j $(w_i, w_j \in D; i < j)$. Finally, Equation 13, presents $CG_m(D)$ (Coherence Global) the average similarities/distances between the embedding of $\tilde{w}_i$ of each word $w_i \in D$ and the pooling of the embeddings of all the words of D, for example their average $avg(\tilde{D})$, because this pooling strategy gives the best results in preliminary experiments.

$$CAW_m(D) = \frac{1}{k-1} \sum_{i=1}^{k-1} m(\tilde{w}_i, \tilde{w}_{i+1}) \quad (11)$$

$$CWP_m(D) = \frac{1}{(k^2-k)/2} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} m(\tilde{w}_i, \tilde{w}_j) \quad (12)$$

$$CG_m(D) = \frac{1}{k} \sum_{i=1}^{k} m(\tilde{w}_i, avg(\tilde{D})) \quad (13)$$

Table 2 shows these coherence measures calculated using the cosine similarity metric (cos) on the changed documents obtained by exchanging the word *"push"* of the original document shown in Figure 3 with words predicted by the BERT MLM. These results are ordered by the word prediction confidence (column %Max. Confidence). We did not find significant correlations between the word prediction rank of BERT MLM and the coherence measures of the CDs derived from the corpora used in our experiments, which are reported in section 5.

Table 2. CDs created by exchanging the word *"push"* in *"How to push elements in a stack ?"* with words predicted by BERT MLM, along with their confidence rank and coherent measures

#	Change Documents (CDs)	%Max. Confidence	CAW_{cos}	CWP_{cos}	CG_{cos}
1	*How to arrange elements in a stack ?*	100.00%	0.653	0.605	0.809
2	*How to organize elements in a stack ?*	96.62%	0.671	0.624	0.819
3	*How to store elements in a stack ?*	50.35%	0.648	0.606	0.809
4	*How to put elements in a stack ?*	38.64%	0.632	0.593	0.803
5	*How to manage elements in a stack ?*	32.72%	0.537	0.508	0.754
6	*How to place elements in a stack ?*	27.79%	0.636	0.594	0.803
7	*How to represent elements in a stack ?*	23.80%	0.638	0.586	0.799
8	*How to stack elements in a stack ?*	23.38%	0.641	0.610	0.811
9	*How to sort elements in a stack ?*	19.21%	0.481	0.481	0.711
10	*How to handle elements in a stack ?*	17.19%	0.650	0.598	0.805

Other alternatives to evaluate the coherence of each document $D \in C(DocSet)$ include, among other possibilities, topic coherence measures like C_{UCI}, C_{NPMI}, C_V and C_{UMass}, which are defined in Section 2.2. For calculating these topic coherence measures, it is necessary to build a topic model using an external corpus from the same domain as the set of documents $DocSet$.

The experiments reported in the following use COHEWL to compare the results of alternative approaches to classify and measure semantic coherence of short documents at the word level.

5 Experimental Setting

COHEWL was implemented in Python version 3.6.9 and has been executed in Google Collaboratory. Document segmentation and POS-tagging were performed using the spaCy tool version 3.2.0. SpaCy leverages a stop word list to filter out common and uninformative words during this process. The full list of stop words used by spaCy can be found in the project's official GitHub repository[4]. POS-Tagging and the stop word list allow selecting certain types of lexicons in datasets when generating contrastive data and in measurement experiments. Coherence classification and measurement with BERT used the Huggingface implementation version 4.5.1[5] for documents in English, and $BERTimbau$ [Souza *et al*., 2020][6] for Portuguese. Both variants were used in the size $Large$ (1,024 dimensions) and the format cased (with uppercase and lowercase characters). The BERT embeddings were handled using PyTorch version 1.8.1. Topic coherence measures were produced using Gensim [Řehůřek and Sojka, 2010] version 4.2.0[7]. The full implementation and results of all experiments are available on Github[8].

5.1 Datasets

Coherence classification and measurement were done for the documents of the 3 datasets described in the following. Table 3 presents for each dataset selected for our experiments its number of documents and averaged numbers per documents of: words; words that are not stop words; words that are nouns, verbs or/and auxiliary verbs; and out of vocabulary words (OOV).

CohQuAD (Coherence Question Answering Dataset)[9] has 40 documents, each one with a single-sentence question in the domain of data structures, inspired from student posts in Virtual Learning Environment (VLE) forums. These documents are made up of sentences of simple or compound periods. For this study, the authors created and manually classified all documents. Half (20) were designed to be coherent, while the other half (20) were crafted to be incoherent. To achieve incoherence, the authors meticulously selected a distinct coherent document and manually substituted a word (verb, noun, or auxiliary verb) with a synonym that alters the meaning (e.g., swapping "queue" for "stack", "stack" for "queue"). Finally, the original Brazilian Portuguese (pt-br) dataset was translated into English (en) using Google Translate, then manually revised to ensure accuracy in both languages for broader experimentation.

FaQuAD [Sayama *et al*., 2019][10] has 900 questions taken from 249 passages of 18 official documents about Computer Science from a Brazilian public university, and 21 Wikipedia articles about University education in Brazil. It follows the SQuAD 1.0 format.

SQuAD 2.0 (Stanford Question Answering Dataset) [Rajpurkar *et al*., 2018][11] has 142,192 documents, each one with a single question in English. A version translated into Portuguese is available at Huggingface Datasets[12]. We excluded from the dataset 266 documents which do not have any verb, auxiliary verb or noun, leaving 141,926 documents in each language. We randomly selected just 1% of the remaining documents (1,419) to make our experiments viable and to balance the dataset size with that of FaQuAD. The selected

[4] https://github.com/explosion/spaCy/blob/master/spacy/lang/en/stop_words.py

[5] https://huggingface.co/docs/transformers/model_doc/bert

[6] https://github.com/neuralmind-ai/portuguese-bert/

[7] https://radimrehurek.com/gensim/

[8] https://github.com/osmarbraz/cohewl/

[9] https://github.com/osmarbraz/cohewl/tree/main/dataset/cohquad/

[10] https://github.com/liafacom/faquad/

[11] https://rajpurkar.github.io/SQuAD-explorer/

[12] https://huggingface.co/datasets/piEsposito/squad_20_ptbr

Table 3. Datasets used in our coherence classification and measurement experiments

	CoQuAD		FaQuAD	SQuAD 2.0	
Measures	(pt-br)	(en)	(pt-br)	(pt-br)	(en)
#Documents	40	40	900	1,419	1,419
Avg #words per document	10.85	11.30	10.82	11.41	11.19
Avg(#words-#stopwords)	5.20	5.00	5.33	6.13	6.16
Avg(#nouns+#verbs+#aux)	4.35	4.60	4.16	3.86	4.11
Avg(#OOV words)	1.90	0.50	1.22	1.32	0.24

documents for the versions in English and Portuguese refer to the same questions in the respective languages.

The construction of the topic model necessary to calculate topic coherence measures used a domain-specific corpus that we call **Stack&Queue**.[13] This corpus, created by the authors, includes text passages about the data structures *stack* and *queue*, taken from Wikipedia and the book *"Introduction to Algorithms"* of Cormen *et al*. [2009]. Versions in English and Brazilian Portuguese were created with original text in the respective languages. Table 4 presents the number of sentences, words, words that are not stop words and the vocabulary size of this corpus in each language.

Table 4. Domain specific corpus used to build the topic model

Measures	Stack&Queue (pt-br)	Stack&Queue (en)
#Sentences	131	109
#Words	2,102	1,996
#Words - #stopwords	1,243	1,114
#Vocabulary	615	519

5.2 Contrastive Data

We generated contrastive data, i.e., documents with a word changed (CDs), for each original document (OD) of the selected datasets. Our experiments used alternatives for the number of modified versions generated per OD, namely $N_c \in \{1, 20, 100\}$, while Barzilay and Lapata [2008] only considered $N_c = 20$. For instance, from the 20 coherent CohQuaD documents we generated 20, 400, and 2,000 modified documents (for N_c equal to 1, 20 and 100, respectively). The data extended with the discriminative CDs were organized in pairs $\langle OD, CD \rangle$, i.e., with each original document pairing with each one of its modified versions. The 5 extended datasets used in the experiments are described below. Notice that only 1 (CohQuAD Coh+Inc) does not include documents with a word substituted by an automatically predicted one. The total number of instances of each extended dataset with contrastive data can be calculated by multiplying the number of documents presented in Table 3 by the values of $N_c \in \{1, 20, 100\}$.

1. CohQuAD Coh+Inc - only human-made and manually classified documents (20 coherent and 20 incoherent);
2. C(CohQuAD Coh) - CohQuAD coherent documents and their versions automatically modified by changing a word;
3. C(CohQuAD Inc) - CohQuAD incoherent documents and their versions automatically modified by changing a word;
4. C(CohQuAD Coh+Inc) - CohQuAD coherent + incoherent documents and their versions automatically modified by changing a word;
5. C(FaQuAD) and C(SQuAD 2.0) - original documents from the respective dataset, their versions automatically modified by changing a word.

[13] https://github.com/osmarbraz/cohewl/tree/main/dataset/stack_queue/

6 Discriminating Original from Changed Documents

These experiments aim to evaluate the performance of our fine-tuned BERT model to distinguish coherent documents from incoherent ones and original documents from the ones with a word automatically changed. The fine-tuning for classifying documents as original (OD) or changed (CD) used the learning rate scheduler without warm-up and with linear decay of the learning rate along the training stages.

We used the AdamW optimizer from the Huggingface implementation of BERT with $\beta_1 = 0.9$, $\beta_2 = 0.999$ (default value). Training was interrupted after a given number of epochs. Hyperparameters were set for each dataset. In total, we run 120 experiments for the distinct datasets and their modified versions to find the optimal classifier. The parameter value ranges combinations were those suggested by Devlin *et al*. [2019]: learning rates of 10^{-5}, $2 * 10^{-5}$, $3*10^{-5}$, $4*10^{-5}$, $5*10^{-5}$ (we included 10^{-5} and $4*10^{-5}$), number of epochs in {2, 3, 4}, batch sizes of 16 and 32 for training and validation, respectively. The classifiers performance was evaluated by averaging accuracies (10-fold cross-validation). We realized only evaluation (no test) due to the limited size of our unique dataset with ground true for word coherence classification (e.g., CoQuAD).

6.1 Classifier Results

Table 5 shows the averaged accuracy (*Acc.*) for 10-fold validation of the binary classifiers (built on $BERTimbau_{Large}$ and $BERT_{Large}$ for documents in Portuguese and English, respectively) per dataset and respective: numbers of modified document versions (N_c), batch sizes (*Bat.*), learning rates(*Lear. rate*), and numbers of epochs (*Ep.*). The best results for each dataset are highlighted in bold. The first two

result lines in this table show hyperparameters that provided that best accuracy for discriminating coherent from incoherent documents of our manually built annotated dataset. The remaining result lines show the hyperparameters that allowed the best accuracy for distinguishing original documents from the ones with a word automatically changed by using BERT MLM. Notice that the accuracy increases with the number of modified versions per OD. For $N_c = 1$ the accuracies of all classifiers surpassed 60.26%, while for $N_c = 20$ the minimum accuracy was 92.55%, and for $N_c = 100$ it was 97.38%. Accuracies for CohQuAD were the highest regardless of the document coherence: coherent (Coh), Incoherent (Inc) or both (Coh+Inc). Figure 4 shows the confusion matrices of the classifiers for each dataset.

Table 5. Hyperparameter configuration and accuracy of the fine-tuned BERT classifier

Extended datasets	N_c	*Bat.*	*Learn. rate*	*Ep.*	*Acc.*
CohQuAD Coh+Inc (pt-br)	0	16	$5 * 10^{-5}$	4	**87.50%**
CohQuAD Coh+Inc (en)	0	16	$4 * 10^{-5}$	4	**80.00%**
C(CohQuAD Coh) (pt-br)	1	16	10^{-5}	3	80.00%
	20	16	10^{-5}	4	97.88%
	100	32	10^{-5}	4	**99.88%**
C(CohQuAD Coh) (en)	1	16	$2 * 10^{-5}$	4	75.00%
	20	16	$3 * 10^{-5}$	4	95.13%
	100	32	10^{-5}	4	**99.80%**
C(CohQuAD Inc) (pt-br)	1	16	$2 * 10^{-5}$	4	82.50%
	20	16	$2 * 10^{-5}$	4	97.50%
	100	16	$2 * 10^{-5}$	4	**99.88%**
C(CohQuAD Inc) (en)	1	16	$5 * 10^{-5}$	4	72.50%
	20	16	$2 * 10^{-5}$	4	96.38%
	100	16	10^{-5}	3	**99.85%**
C(CohQuAD Coh+Inc) (pt-br)	1	16	$3 * 10^{-5}$	4	87.50%
	20	16	10^{-5}	4	99.31%
	100	32	10^{-5}	3	**99.79%**
C(CohQuAD Coh+Inc) (en)	1	16	$5 * 10^{-5}$	4	87.50%
	20	16	10^{-5}	4	99.44%
	100	32	10^{-5}	4	**99.83%**
C(FaQuAD) (pt-br)	1	16	$2 * 10^{-5}$	3	64.89%
	20	16	10^{-5}	4	95.65%
	100	16	10^{-5}	3	**98.76%**
C(SQuAD 2.0) (pt-br)	1	16	$3 * 10^{-5}$	4	64.59%
	20	16	10^{-5}	4	95.69%
	100	16	10^{-5}	3	**98.90%**
C(SQuAD 2.0) (en)	1	16	$2 * 10^{-5}$	4	60.26%
	20	32	10^{-5}	4	92.55%
	100	16	10^{-5}	**4**	**97.38%**

6.2 Discussion of the Classifier Results

The high accuracy of the classifiers can be explained by analyzing the distribution of the document embeddings produced by the pooling strategy proposed in Reimers and Gurevych [2019], that uses the average of word embeddings. The Embedding Projector [Smilkov *et al.*, 2016] was used to

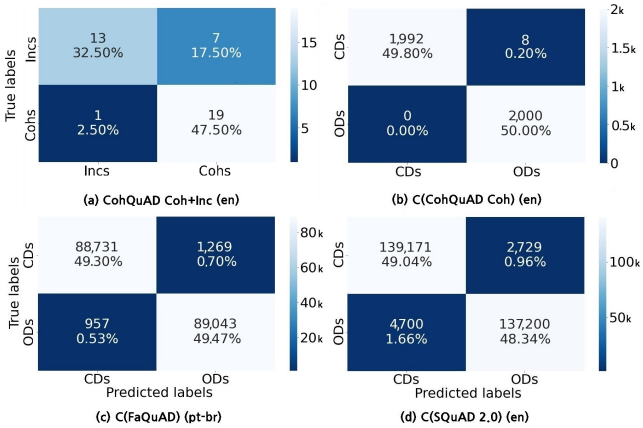

Figure 4. Confusion matrix of the best classification results

reduce embedding dimensionality and visualize their projection in a plane. We chose to make these projections using t-Distributed Stochastic Neighbor Embedding (t-SNE), a popular non-linear dimensionality reduction technique [Van der Maaten and Hinton, 2008], because usually preserves some local structure (clusters) which allows inspecting the nearest neighbors of a given point [Smilkov *et al.*, 2016].

Figure 5 presents two-dimensional t-SNE projections of the sentence embeddings from the datasets: C(CohQuAD Coh) and C(SQuAD 2.0) in Brazilian Portuguese and English. All projections were generated with 250 iterations over 20 original documents plus their 20 modified versions, resulting in a total of 420 points per projection. Just 20 random documents from SQuaD 2.0 were selected for image clarity. In all projections, each red dot represens an OD and each blue dot a CD. Notice that CD versions of the same OD form clusters in all projections. CD versions tend to be close to the respective OD. Each OD is usually slightly separated from the respective CD versions in the vast majority of clusters. In projection (a), some red dots are close or overlapping due to the high similarity of these documents. These distributions probably contributed to the high classification accuracies, above 92% for Nc equal to 20 and 100.

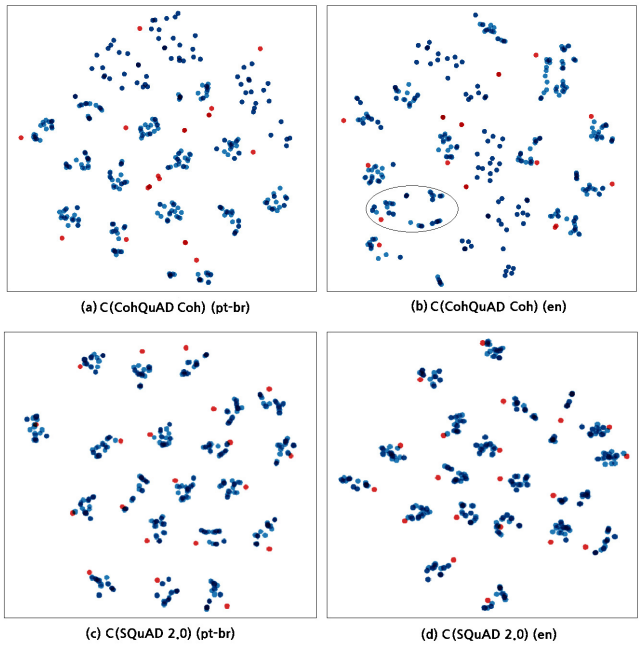

Figure 5. t-SNE projections of document embeddings (OD red, CD blue)

7 Measuring Semantic Coherence with COHEWL

These experiments aim to evaluate the effectiveness of the alternative formulas presented in subsection 4.2 for measuring coherence. They investigate how the (in)coherence measures differ for coherent documents, incoherent documents and documents with a word automatically changed by using the BERT MLM.

The measures of (in)coherence $C_m \in \{CAW_m, CWP_m, CG_m\}$ were calculated in accordance with Equation 11, 12 and 13, using the similarity and distance measures $m \in \{cos, euc, man\}$, i.e., cosine similarity (cos), Euclidean (euc) and Manhattan (man) distances. We employed two word pooling strategies, namely *MEAN* and *MAX*, for out-of-vocabulary word tokens and three alternatives for selecting words from short texts: all words (*ALL*), removal of stop words (*CLEAN*) and only relevant words from the morpho-syntactic classes verb, noun or auxiliary verb (*REL*). Combining these alternatives, 54 experiments were performed per dataset to calculate measures for each original document (OD) and their $N_c \in \{1, 20, 100\}$ modified versions (CD).

7.1 (In)coherence Measurement Results

Table 6 presents the percentages of pairs $\langle OD, CD \rangle$ (each OD paired with N_c CDs) for which the (in)coherence measure C_m of the OD is better than that of the respective CD in the pair, for distinct values of N_c, pooling strategies (*Pool.*), word selection alternatives $Word\,sel. \in \{ALL, CLEAN, REL\}$, (in)coherence measures $C \in \{CAW, CWP, CG\}$, and distance/similarity functions $m \in \{euc, man, cos\}$. The first two result lines in this table show the strategies and measures that best distinguished coherent from incoherent documents of our manually built annotated dataset. The remaining result lines show the strategies and measures that best distinguished original documents with the ones with a word changed by using BERT MLM.

The graphics in Figure 6 show, in ascending order, the coherence measures CAW_{cos} (red), CWP_{cos} (blue), and CG_{cos} (green) of each original document OD from C(CohQuAD Coh) and C(SQuAD 2.0) in Brazilian Portuguese (pt-br) and English (en). The (in)coherence measures (represented by dashed lines) of the 20 modified versions of CD are in ascending order for the respective OD. We do not present C(FaQuAD) results because they have the same characteristics as C(SQuAD 2.0). The highest measure gains for modified versions CD over an OD are highlighted with an enclosing red rectangle on the left of the graphic in Figure 6(b).

Figure 7 shows the coherence measures of the OD (dashed lines) from C(CohQuAD Coh) (en): *"How to push elements in a stack ?"* (highlighted in Figures 5(b) and 6(b)), those of its CDs (dashed lines) and the BERT MLM confidence for each prediction for the masked word *push* used to change the OD into a CD.

7.2 Discussion of the (In)coherence Measurement Results

Figure 8 presents the PCA projection of word and sentence embeddings derived from the C(CohQuAD Coh) (en) OD *"How to push elements in a stack?"* and the first word (*arrange*) predicted by BERT MLM to replace *push*. Blue dots represent words from the original document OD, red dots represent words from the changed document CD, the cyan dot signifies the OD itself, and the pink dot marks the CD.

Contextualized embeddings for words with similar meanings or functions (e.g., "How", "to", "?") tend to cluster together, regardless of whether they appear in the OD or CD. However, these embeddings do not consistently reflect human perception of text coherence. For instance, the distance between "push" and "stack" in the embedding space does not align with human perception of their close semantic relationship in the domain of of data structures . Unfortunately, in the embedded space, the distance between "push" and "stack" is slightly bigger than that between"push" and "arrange". Therefore, one cannot rely on the word distances in the embedded space to capture the word coherence perceived by humans.

8 Topic Coherence Measures

These experiments aim to evaluate the effectiveness of some topic coherence measures for measuring document coherence. The topical coherence measures $TC \in \{C_{\text{UCI}}, C_{\text{NPMI}}, C_V, C_{\text{UMass}}\}$ were calculated as specified in Equations 2, 4, 6 and 9, respectively, for each document in the CohQuAD Coh+Inc dataset (en and pt-br). Three word selection alternatives (*ALL*, *CLEAN*, *REL*) and sliding window sizes (sw) between 3 and 150 were used. Combining these alternatives, a total of 444 experiments were performed. The CohQuAD Coh+Inc dataset and the Stack&Queue corpus used to build the topic model were pre-processed to remove punctuation marks (e.g.",","?") and lowercased all characters. POS-Tagging was employed to select only verbs and nouns for calculating coherence measures in the word selection alternative *REL*.

8.1 Topic Coherence Measurement Results

The last column of Table 7 presents the percentages of $\langle OD, CD \rangle$ pairs from CohQuAD Coh+Inc dataset (en and pt-br) for which the topical coherence measure of OD is higher than that of CD. These results are detailed for distinct alternatives to select the words (*Word select*) from each document to be considered in the measure calculations, the sliding *Window size* and the topic coherence measure (*TC*). The best results for each dataset are in bold.

Figure 9 details the distribution of the topical coherence measures among CohQuAD Coh+Inc (en) documents, considering only the parameter settings that maximized the measure differences between each original document OD and its respective changed documents CD. Solid lines refer to measures of OD, while the dashed lines CD. Notice that the average distance between the solid line (OD) and the dashed

Table 6. Proportion of pairs $\langle OD, CD \rangle$ with $C_m(OD) > C_m(CD)$

Datasets	N_c	$Pool.$	$Word\,sel.$	C	m	%
CohQuAD Coh+Inc (pt-br)	**0**	**MEAN**	**CLEAN**	**CWP**	**man**	**65.00%**
CohQuAD Coh+Inc (en)	**0**	**MAX**	**ALL**	**CWP**	**euc**	**60.00%**
C(CohQuAD Coh) (pt-br)	**1**	**MEAN**	**ALL**	**CAW**	**man**	**85.00%**
	20	MEAN	ALL	CAW	man	80.25%
	100	MEAN	ALL	CAW	man	80.70%
C(CohQuAD Coh) (en)	**1**	**MEAN**	**REL**	**CAW**	**man**	**60.00%**
	20	MEAN	ALL	CWP	euc	49.50%
	100	MEAN	ALL	CWP	euc	46.40%
C(CohQuAD Inc) (pt-br)	1	MEAN	ALL	CAW	cos	65.00%
	20	**MEAN**	**ALL**	**CAW**	**man**	**77.25%**
	100	MEAN	CLEAN	CG	cos	72.80%
C(CohQuAD Inc) (en)	**1**	**MEAN**	**REL**	**CWP**	**man**	**75.00%**
	20	MEAN	ALL	CAW	euc	51.25%
	100	MEAN	REL	CWP	man	51.75%
C(CohQuAD Coh+Inc) (pt-br)	1	MEAN	ALL	CAW	man	70.00%
	20	**MEAN**	**ALL**	**CAW**	**man**	**77.88%**
	100	MEAN	ALL	CAW	man	77.08%
C(CohQuAD Coh+Inc) (en)	**1**	**MEAN**	**REL**	**CWP**	**man**	**65.00%**
	20	MEAN	ALL	CWP	euc	51.63%
	100	MEAN	ALL	CWP	euc	50.40%
C(FaQuAD) (pt-br)	1	MEAN	CLEAN	CWP	cos	67.00%
	20	MEAN	REL	CG	euc	60.93%
	100	**MEAN**	**REL**	**CAW**	**euc**	**74.60%**
C(SQuAD 2.0) (pt-br)	1	MEAN	CLEAN	CWP	cos	70.33%
	20	MEAN	CLEAN	CG	euc	61.06%
	100	**MEAN**	**ALL**	**CWP**	**euc**	**71.52%**
C(SQuAD 2.0) (en)	**1**	**MEAN**	**REL**	**CWP**	**cos**	**71.04%**
	20	MEAN	REL	CWP	cos	59.50%
	100	MEAN	REL	CWP	cos	57.57%

line (CD) for each measure (distinct color) is higher for C_{UCI}, C_{NPMI} and C_V than for C_{UMass}. In addition, only document 20 has a topic measure (C_{UMass}) lower for OD than for the respective CD.

Finally, Figure 10 shows the impact of the sliding window size (3-40) in the proportion of OD with topic coherence higher than that of the respective CD in CohQuAD Coh+Inc (en). We used smaller sliding window sizes than the proposed by [Röder *et al*., 2015], because CohQuAD Coh+Inc (en) are short (11.3 words per document in average). The $CLEAN$ scheme was used to select words for calculating the topic coherence measures C_{UCI}, C_{NPMI} and C_V. The C_{UMass} measure was not included because it does not use a sliding window.

8.2 Discussion of the Topic Coherence Measurement Results

Topic coherence measures seem to better capture short documents coherence than measures based on the average of embedding similarities or distances. Our experiments also showed that C_{UCI} and C_{NPMI} produced the best results using the $CLEAN$ alternative to select relevant words, followed by REL and ALL. It suggests that certain words are more relevant for topic coherence.

9 Conclusions and Future Work

This work introduced the COHEWL framework to assess the semantic coherence of short documents based on the semantic compatibility of their words. COHEWL supports contrastive data generation via word replacement, coherence assessment using alternative classification and measuring methods, and performance comparison. The current COHEWL implementation, described and evaluated in this article, used the BERT MLM to predict word substitutes for generating contrastive data (documents with a modified word). BERT was fine-tuned to distinguish original from modified documents. Coherence measures based on average similarities or distances between BERT embeddings or, alternatively, topical coherence measures were also calculated. Their distributions were compared with each other, with the BERT MLM word prediction rank, and with human appreciation of coherence in a manually made dataset of coherent and incoherent documents in the domain of data structures.

Experimental results revealed that, in a similar way as BERT provides high accuracy for the Shuffle Test, it also performs very well in the task of discriminating original short documents from the ones with a word changed. Nevertheless, (in)coherence measures based on the average similarities or distances between embeddings generated by BERT are not adherent to document classification. The use of topic coher-

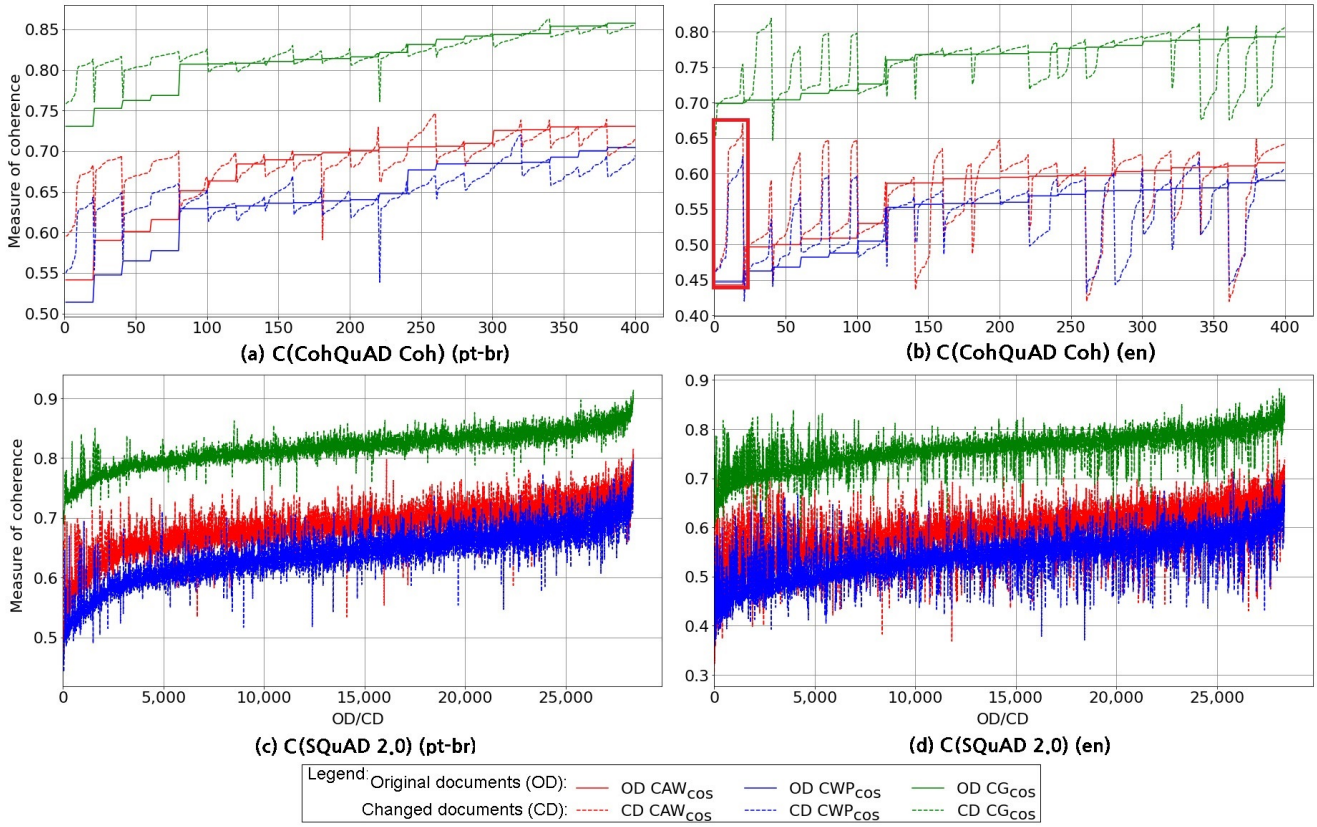

Figure 6. Coherence measures in ascending order

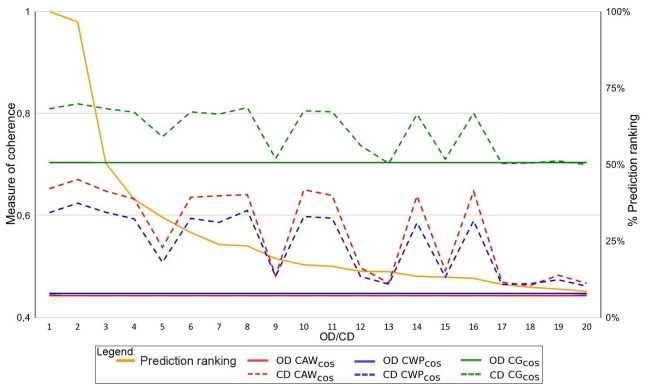

Figure 7. Prediction confidence and coherence measures of the OD *"How to push elements in a stack ?"* and its 20 CDs

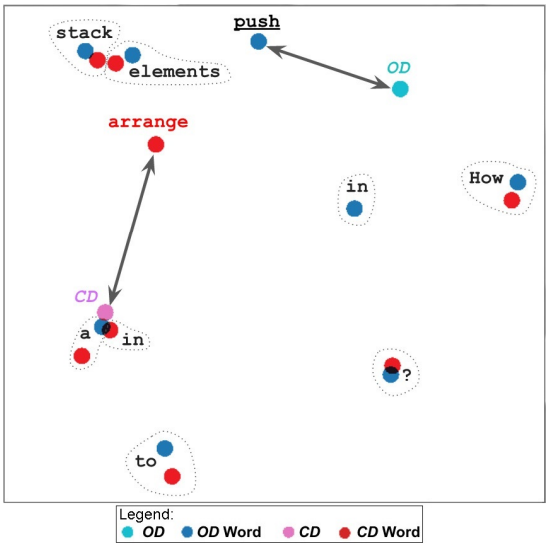

Figure 8. PCA projection of word and sentence embeddings of *"How to push elements in a stack?"* and those of the CD produced by changing the word *"push"* with *"arrange"*

ence measures calculated with a model built from a domain specific small corpus proved more adherent to the human perception of coherence at the level of words.

The results presented in this paper suggest that investigating alternatives to the current language models and embeddings to assess the semantic coherence of short documents at the word level, such as new topic and embedding models like Gao *et al*. [2019]; Churchill and Singh [2022, 2023] can be a promising research direction. Identifying words involved in sometimes subtle semantic incoherences and automatically solving them by predicting more suitable words for each situation are challenging tasks yet. The BERT contextualized embeddings of these words are not necessarily outliers in the embedding spaces, as we have shown in the reported experimental results. Detecting and solving semantic incoherences at the word level can depend on capturing word meanings specific to particular domains. Nevertheless, the methods proposed in this paper allow at least detecting some of these incoherences and can help open new frontiers in the study of semantic coherence. Progresses in this field can benefit a variety of downstream tasks and applications of natural language processing, such as question answering [Bouarroudj *et al*., 2022].

The COHEWL framework presents certain vulnerabilities and threats. One potential risk lies in the reliance on BERT MLM for word replacement, which may generate contrastive data that does not always align with human perception of coherence. Additionally, coherence measures based on em-

Table 7. Proportion of pairs $\langle OD, CD \rangle$ with $TC(OD) > TC(CD)$

Extended datasets	*Word select*	*Window size*	*TC*	%
CohQuAD Coh+Inc (pt-br)	ALL	16	C_{UCI}	95.00%
		17	C_{NPMI}	90.00%
		3	C_V	65.00%
		-	C_{UMass}	85.00%
	CLEAN	10	C_{UCI}	**95.00%**
		5	C_{NPMI}	**95.00%**
		10	C_V	**95.00%**
		-	C_{UMass}	**95.00%**
	REL	9	C_{UCI}	90.00%
		3	C_{NPMI}	85.00%
		11	C_V	80.00%
		-	C_{UMass}	90.00%
CohQuAD Coh+Inc (en)	ALL	9	C_{UCI}	95.00%
		9	C_{NPMI}	95.00%
		39	C_V	80.00%
		-	C_{UMass}	90.00%
	CLEAN	3	C_{UCI}	**100.00%**
		3	C_{NPMI}	**100.00%**
		15	C_V	95.00%
		-	C_{UMass}	**95.00%**
	REL	3	C_{UCI}	95.00%
		3	C_{NPMI}	95.00%
		15	C_V	**100.00%**
		-	C_{UMass}	95.00%

bedding similarities or distances proved inadequate for document classification, indicating a limitation in capturing semantic coherence effectively. The framework's performance is also constrained by domain-specific small corpora, affecting generalization.

Future work related with text coherence and the results presented in this paper include: (*i*) experiments with more varied and larger manually labeled datasets to better analyze the adherence of distinct classifiers and coherence measures to the human perception of text coherence; (*ii*) investigate the impact of features derived from style, syntax, and combinations of word and knowledge embeddings to classify and measure coherence; (*iii*) use prompts in contextualized models such as GPT in the evaluation of coherence; and (*iv*) devise methods for identifying particular words involved in semantic incoherences and automatically exchange them with other ones to produce more coherent documents.

Declarations

Funding

This work was supported by the State University of Santa Catarina (UDESC), a CNPq Universal grant, FAPESC grant 2021TR1510 and the Print CAPES-UFSC Automation 4.0 Project. The Céos project, financed by the Public Ministry of Santa Catarina State (MPSC), has also contributed a lot to improve our infrastructure.

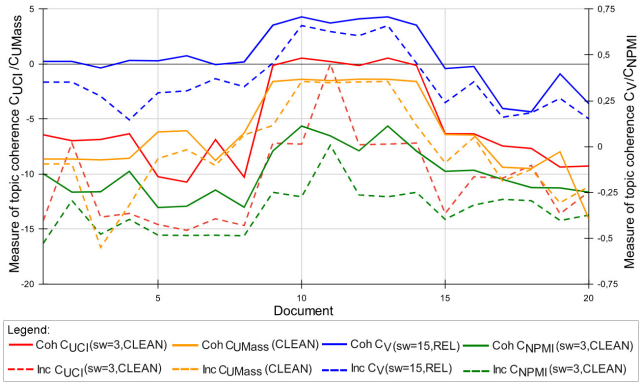

Figure 9. Best topic coherence measures for CohQuAD Coh+Inc (en) documents

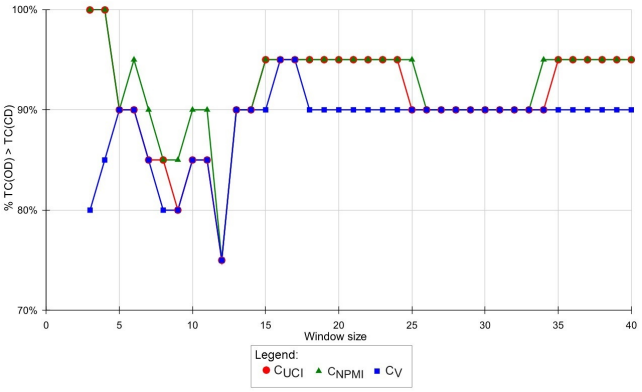

Figure 10. Impact of sliding window size on $\%(TC(OD) > TC(CD))$ in CohQuAD Coh+Inc (en)

Authors' Contributions

Osmar de Oliveira Braz Junior: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review and editing, Visualization. **Renato Fileto**: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Writing – original draft, Writing – review and editing, Visualization, Supervision, Project administration.

Competing interests

The authors declare that they have no conflict of interest.

Availability of data and materials

Materials are available on the links provided in the article. In case the links become broken, the materials can be made available upon request.

References

Aletras, N. and Stevenson, M. (2013). Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, pages 13–22, Potsdam, Germany. Association for Computational Linguistics. Available at:`https://aclanthology.org/W13-0102.pdf`.

Bao, M., Li, J., Zhang, J., Peng, H., and Liu, X. (2019). Learning semantic coherence for machine generated spam

text detection. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, Budapest, Hungary. IEEE. DOI: 10.1109/IJCNN.2019.8852340.

Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34. DOI: 10.1162/coli.2008.34.1.1.

Barzilay, R. and Lee, L. (2004). 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*, pages 113–120, Boston, Massachusetts, USA. Association for Computational Linguistics. DOI: 10.48550/arXiv.cs/0405039.

Bouarroudj, W., Boufaida, Z., and Bellatreche, L. (2022). Named entity disambiguation in short texts over knowledge graphs. *Knowledge and Information Systems*, 64(2):325–351. DOI: 10.1007/s10115-021-01642-9.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of German Society for Computational Linguistics and Language Technology (GSCL 2009)*, 30:31–40. Available at:`https://svn.spraakdata.gu.se/repos/gerlof/pub/www/Docs/npmi-pfd.pdf`.

Braz Jr., O. O. and Fileto, R. (2021). Investigando coerência em postagens de um fórum de dúvidas em ambiente virtual de aprendizagem com o BERT. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, pages 749–759. SBC. DOI: 10.5753/sbie.2021.217397.

Church, K. W. and Hanks, P. (1989). Word association norms, mutual information, and lexicography. In *Proceedings of the 27th Annual Meeting on Association for Computational Linguistics*, ACL '89, page 76–83, USA. Association for Computational Linguistics. DOI: 10.3115/981623.981633.

Churchill, R. and Singh, L. (2022). The evolution of topic modeling. *ACM Computing Surveys*, 54(10s):1–35. DOI: 10.1145/3507900.

Churchill, R. and Singh, L. (2023). Using topic-noise models to generate domain-specific topics across data sources. *Knowledge and Information Systems*, pages 1–28. DOI: 10.1007/s10115-022-01805-2.

Cormen, T. H., Stein, C., Rivest, R. L., and Leiserson, C. E. (2009). *Introduction to Algorithms*. MIT Press, New York, 3nd edition. Book.

Das, D. (2019). Nuclearity in RST and signals of coherence relations. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 30–37, Minneapolis, MN. Association for Computational Linguistics. DOI: 10.18653/v1/W19-2705.

De Beaugrande, R.-A. and Dressler, W. U. (1981). *Introduction to Text Linguistics*, volume 1. Longman, London. Available at:`http://library.lgaki.info:404/2017/De%20Beaugrande%20R.-A..pdf`.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. DOI: 10.18653/v1/N19-1423.

Farag, Y. and Yannakoudakis, H. (2019). Multi-task learning for coherence modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 629–639, Florence, Italy. Association for Computational Linguistics. DOI: 10.18653/v1/P19-1060.

Foltz, P. W., Kintsch, W., and Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307. DOI: 10.1080/01638539809545029.

Gao, W., Peng, M., Wang, H., Zhang, Y., Xie, Q., and Tian, G. (2019). Incorporating word embeddings into topic modeling of short text. *Knowledge and Information Systems*, 61:1123–1145. DOI: 10.1007/s10115-018-1314-7.

Halliday, M. A. K. and Hasan, R. (2014). *Cohesion in English*. English language series. Taylor & Francis, London. Book.

Joty, S., Mohiuddin, M. T., and Tien Nguyen, D. (2018). Coherence modeling of asynchronous conversations: A neural entity grid approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 558–568, Melbourne, Australia. Association for Computational Linguistics. DOI: 10.18653/v1/P18-1052.

Koch, I. G. V. and Travaglia, L. C. (2021). *A coerência textual*. Repensando a língua portuguesa. Editora Contexto, São Paulo, Brasil. Available at: `https://dialogo.fflch.usp.br/sites/dialogo.fflch.usp.br/files/upload/paginas/KOCH%2C%20I.%20A%20COERENCIA%20TEXTUAL%20-%20co%CC%81pia_0.pdf`.

Laban, P., Dai, L., Bandarkar, L., and Hearst, M. A. (2021). Can transformer models measure coherence in text: Rethinking the shuffle test. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1058–1064, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2021.acl-short.134.

Li, J. and Jurafsky, D. (2017). Neural net models of open-domain discourse coherence. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 198–209, Copenhagen, Denmark. Association for Computational Linguistics. DOI: 10.18653/v1/D17-1019.

Lin, Z., Ng, H. T., and Kan, M.-Y. (2011). Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 997–1006, Portland, Oregon, USA. Association for Computational Linguistics. Available at:`https://aclanthology.org/P11-1100.pdf`.

Mesgar, M. and Strube, M. (2018). A neural local coherence model for text quality assessment. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4328–4339, Brussels, Belgium. Association for Computational Linguistics. DOI:

10.18653/v1/D18-1464.

Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 262–272, Edinburgh, Scotland, UK. Association for Computational Linguistics. Available at:https://aclanthology.org/D11-1024.Pdf.

Mohiuddin, M. T., Jwalapuram, P., Lin, X., and Joty, S. (2021). Rethinking coherence modeling: Synthetic vs. downstream tasks. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3528–3539, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2021.eacl-main.308.

Moon, H. C., Mohiuddin, T., Joty, S., and Xu, C. (2019). A unified neural coherence model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2262–2272, Hong Kong, China. Association for Computational Linguistics. DOI: 10.18653/v1/D19-1231.

Muller, P., Braud, C., and Morey, M. (2019). ToNy: Contextual embeddings for accurate multilingual discourse segmentation of full documents. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 115–124, Minneapolis, MN. Association for Computational Linguistics. DOI: 10.18653/v1/W19-2715.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 100–108, USA. Association for Computational Linguistics. Available at:https://aclanthology.org/N10-1012.pdf.

Nie, A., Bennett, E., and Goodman, N. (2019). DisSent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510, Florence, Italy. Association for Computational Linguistics. DOI: 10.18653/v1/P19-1442.

Rajpurkar, P., Jia, R., and Liang, P. (2018). Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics. DOI: 10.18653/v1/P18-2124.

Řehůřek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. Available at:https://radimrehurek.com/lrec2010_final.pdf.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics. DOI: 10.18653/v1/D19-1410.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, WSDM '15, pages 399–408, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/2684822.2685324.

Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M., and Okruszek, L. (2021). Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135. DOI: 10.1016/j.psychres.2021.114135.

Sayama, H. F., Araujo, A. V., and Fernandes, E. R. (2019). Faquad: Reading comprehension dataset in the domain of brazilian higher education. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 443–448. IEEE. DOI: 10.1109/BRACIS.2019.00084.

Shen, A., Mistica, M., Salehi, B., Li, H., Baldwin, T., and Qi, J. (2021). Evaluating document coherence modeling. *Transactions of the Association for Computational Linguistics*, 9:621–640. DOI: 10.1162/tacl_a_00388.

Smilkov, D., Thorat, N., Nicholson, C., Reif, E., Viégas, F. B., and Wattenberg, M. (2016). Embedding projector: Interactive visualization and interpretation of embeddings. *CoRR*, abs/1611.05469. DOI: 10.48550/arxiv.1611.05469.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Brazilian Conference on Intelligent Systems*, pages 403–417, Rio Grande, Brazil. Springer, Springer. DOI: 10.1007/978-3-030-61377-8_28.

Vakulenko, S., de Rijke, M., Cochez, M., Savenkov, V., and Polleres, A. (2018). Measuring semantic coherence of a conversation. In *The Semantic Web – ISWC 2018*, pages 634–651, Cham. Springer International Publishing. DOI: 10.1007/978-3-030-00671-6_37.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(86):2579–2605. Available at:https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf?fbcl.

Van Dijk, T. A., Kintsch, W., *et al*. (1983). *Strategies of discourse comprehension*. Academic Press New York, New York, New York. Available at:https://discourses.org/wp-content/uploads/2022/06/Teun-A-van-Dijk-Walter-Kintsch-1983-Strategies-Of-Discourse-Comprehension.pdf.

Wadud, M. A. H. and Rakib, M. R. H. (2021). Text coherence analysis based on misspelling oblivious word embeddings and deep neural network. *International Journal of Advanced Computer Science and Applications*, 12(1). DOI: 10.14569/IJACSA.2021.0120124.

Wang, W., Feng, S., Wang, D., and Zhang, Y. (2019). Answer-guided and semantic coherent question generation in open-domain conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5066–5076, Hong Kong, China. Association for Compu-

tational Linguistics. DOI: 10.18653/v1/D19-1511.

Wang, Y. and Guo, M. (2014). A short analysis of discourse coherence. *Journal of Language Teaching and Research*, 5(2):460. DOI: 10.4304/jltr.5.2.460-465.

Xu, P., Saghir, H., Kang, J. S., Long, T., Bose, A. J., Cao, Y., and Cheung, J. C. K. (2019). A cross-domain transferable neural coherence model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 678–687, Florence, Italy. Association for Computational Linguistics. DOI: 10.18653/v1/P19-1067.

Zou, W. Y., Socher, R., Cer, D., and Manning, C. D. (2013). Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398, Seattle, Washington, USA. Association for Computational Linguistics. Available at:`https://aclanthology.org/D13-1141.pdf`.