Entity-based Neural Local Coherence Modeling

Anonymous ACL submission

Abstract

In this paper, we propose an entity-based neural local coherence model which is linguistically more sound than previously proposed neural coherence models. Recent neural coherence models encode the input document using large-scale pretrained language models. Hence their basis for computing local coherence are words and even sub-words. The analysis of their output shows that these models frequently compute coherence on the basis of connections between (sub-)words which, from a linguistic perspective, should not play a role. Still, these models achieve state-of-the-art performance in several end applications. In contrast to these models, we compute coherence on the basis of entities by constraining the input to noun phrases and proper names. This provides us with an explicit representation of the most important items in sentences leading to the notion of focus. This brings our model linguistically in line with pre-neural models of computing coherence. It also gives us better insight into the behaviour of the model thus leading to better explainability. Our approach is also in accord with a recent study (O'Connor and Andreas, 2021), which shows that most usable information is captured by nouns and verbs in transformer-based language models. We evaluate our model on three downstream tasks showing that it is not only linguistically more sound than previous models but also that it outperforms them in end applications[1].

1 Introduction

Coherence describes the semantic relation between elements of a text. It recognizes how well a text is organized to convey the information to the reader effectively. Modeling coherence can be beneficial to any system which needs to process a text.

Recent neural coherence models (Mesgar and Strube, 2018; Moon et al., 2019) encode the input document using large-scale pretrained language models (Peters et al., 2018). These neural models compute local coherence, semantic relations between items in adjacent sentences, on the basis of words and even sub-words.

However, it has been unclear on which basis these models compute local coherence. Jeon and Strube (2020) present a neural coherence model, which allows to interpret focus information for the first time. Their investigation reveals that neural models, adopting large-scale pretrained language models, frequently compute coherence on the basis of connections between any (sub-)words or function words. In these cases, the model might capture the focus based on spurious information. While such a model might reach or set the state of the art in some end applications, it will do so for the wrong reasons from a linguistic perspective.

This problem did not appear with pre-neural models of coherence, since they compute coherence on the basis of entities. Early work about pronoun and anaphora resolution by Sidner (1981, 1983) assumes that there is one single salient entity in a sentence, its focus, which serves as a preferred antecedent for anaphoric expressions. Centering theory (Joshi and Weinstein, 1981; Grosz et al., 1995) builds on these insights and introduces an algorithm for tracking changes in focus. Centering theory serves as basis for many researchers to develop systems computing local coherence based on the approximations of entities (Barzilay and Lapata 2008; Feng and Hirst 2012; Guinaudeau and Strube 2013, inter alia).

In this paper, we propose a neural coherence model which is linguistically more sound than previously proposed neural coherence models. We compute coherence on the basis of entities by constraining our model to capture focus on noun phrases and proper names. This provides us with an explicit representation of the most important items in sentences, leading to the notion of focus.

[1] Our implementation will be publicly available upon publication.

This brings our model linguistically in line with pre-neural models of coherence.

Our approach is not only linguistically more sound but also is in accord with the recent empirical study by O'Connor and Andreas (2021) who investigate what contextual information contributes to accurate predictions in transformer-based language models. Their experiments show that most usable information is captured by nouns and verbs. Their findings suggest that we can design better neural models by focusing on specific context words. Our work follows their findings by modeling entity-based coherence in an end-to-end framework to improve a neural coherence model.

Our model integrates a local coherence module with a component which takes context into account. Our model first encodes a document using a pre-trained language model and identifies entities using an linguistic parser. The local coherence module captures the most related representations of entities between adjacent sentences, the local focus. Then it tracks the changes of local foci. The second component captures the context of a text by averaging sentence representations.

We evaluate our model on three downstream tasks: automated essay scoring (AES), assessing writing quality (AWQ), and assessing discourse coherence (ADC). AES and AWQ determine text quality for a given text, aiming to replicate human scoring results. Since coherence is an essential factor in assessing text quality, many previous coherence models are evaluated on AES and AWQ. ADC evaluates coherence models on informal texts such as emails and online reviews. In our evaluation, our model achieves state-of-the-art performance.

We also perform a series of analyses to investigate how our model works. Our analyses show that capturing focus on entities gives us better insight into the behaviour of the model, leading to better explainability. Using this information, we examine the statistical differences of texts assigned to different qualities. From the perspective of local coherence, we find that texts of higher quality are neither semantically too consistent nor too variant. Finally, we inspect error cases to investigate how the models achieve their performance differently.

2 Related Work

Entity-based modeling has been the prevailing approach to model coherence in pre-neural models. The entity grid is its most well-known implementation (Barzilay and Lapata, 2008). It represents entities in a two-dimensional array to track their transitions between sentences. Many variations have been proposed to improve this model, e.g., projecting the grid into a graph representation (Guinaudeau and Strube, 2013) or converting the grid to a neural model (Tien Nguyen and Joty, 2017).

However, the neural version of the entity grid (Tien Nguyen and Joty, 2017) has two limitations. First, Lai and Tetreault (2018) state that entity grids applied to downstream tasks are often extremely sparse. In their evaluation, it is difficult to find meaningful entity transitions between sentences in the grids. Accordingly, this model performs worse than other neural models. More importantly, this neural model cannot provide any clues of how this model works since Tien Nguyen and Joty (2017) apply a convolutional layer on the entity grid. The feature map of the convolutional layer is not interpretable. They cannot examine which entity is assigned more importantly than others by their model. In contrast, we constrain our model to capture focus on entities using noun phrases. Then our model tracks the changes of focus. Hence, it provides us with an interpretable focus (Section 5).

More recently, Moon et al. (2019) propose a neural coherence model to exploit both local and structural aspects. They evaluate their model on an artificial task only, the shuffle test, which determines whether sentences in a document are shuffled or not. However, recent studies (Pishdad et al., 2020) claim that this artificial task is not suitable to evaluate coherence models. Lai and Tetreault (2018) show that the neural coherence models, which achieve the best performance on this task, do not outperform non-neural models on downstream tasks. More recently, Mohiuddin et al. (2021) find a weak correlation between the model performance in artificial tasks and downstream tasks. In our evaluation, we compare Moon et al. (2019) with ours in an artificial task as well as in three downstream tasks. Moon et al. (2019) perform the best in the artificial task, but do not outperform our model in three downstream tasks (Section 4).

3 Our Model

Figure 1 presents an overview of our model architecture. We first introduce our entity representation and sentence encoding using a pretrained language model. Next, we describe a novel local coherence model. We then combine the two representations of

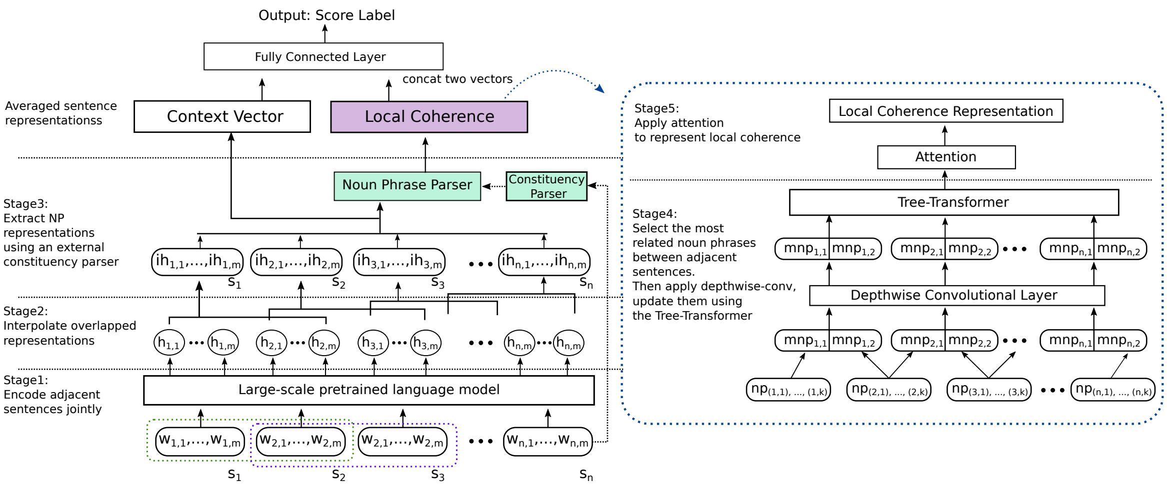


Figure 1: Our model architecture.

local coherence and the context vector, simply averaged sentence representations. Finally, we apply a feedforward network to produce a score label.

3.1 Entity Identification

Pretrained language models encode sequences as sub-words, but to our knowledge, there is no linguistic parser using sub-words as input. Hence, we use an linguistic parser to identify noun phrases in each sentence separately. Kitaev and Klein (2018) present a neural constituency parser which determines the syntactic structure of a sentence. To identify noun phrases and proper names, we apply this parser to the original sentences, then map parsed constituents to sub-word tokens.

Since pretrained language models do not have the means to represent phrase meaning composition, we average sub-word representations for phrases which consists of multiple sub-words. While this implementation does not capture the complicate meaning of phrases, Yu and Ettinger (2020) report that it shows higher correlation with human annotations than using the last word of phrases, assuming that the last word of a phrase is its head.

3.2 Sentence Encoding

We use a pretrained language model (Yang et al., 2019) to encode sentences. XLNet learns bidirectional contexts by maximizing expected likelihood using an autoregressive training objective, hence it has the advantage of capturing focus in sentences. XLNet outperforms other language models in tasks which require to process long texts.

Recent work investigates that pretrained language models learn linguistic features that are helpful for language understanding (Tenney et al., 2019; Warstadt et al., 2020). Inspired by this, we encode two adjacent sentences at once to capture discourse features, such as coreference relations. In this strategy, items are encoded twice except the items included in the first and the last sentence. We interpolate items encoded twice to consider context with regard to the preceding and succeeding sentence.

We first encode an input document using XLNet to obtain word representations. Sentence representations are means of all word representations in a sentence. We then feed sentence representations and the noun phrase representations into the coherence modules.

3.3 Local Coherence Module

We compare the semantic representations of noun phrases between adjacent sentences. The two most similar representations of noun phrases are determined as local focus of the respective sentences. These two representations are averaged to capture the common context. We use cosine similarity to measure semantic similarity.

We notice that some sentences do not include noun phrases, approximately 3.5% in the three datasets used in our evaluation. This mostly occurs when some words are omitted as in cases of ellipsis (Hardt and Romero, 2004). In such cases, we maintain the focus of the previous sentence to preserve the context.

A depthwise convolutional layer is applied to the local focus to record its transitions. Unlike a typical convolutional layer, the depthwise convolutional layer captures the patterns of semantic changes between different time-steps for the same spatial information (Chollet, 2017). Hence, this layer captures the semantic changes between local foci considering the context, but it does not hurt the explainability of our model. We use the lightweight depthwise convolutional layer (Wu et al., 2019).

Then we update the representations of local foci to track the semantic changes between them. We use the Tree-Transformer which updates its hidden representations by inducing a tree-structure from a document (Wang et al., 2019). It generates constituent priors by calculating neighboring attention which represents the probability of whether adjacent items are in the same constituent. The constituent priors constrain the self-attention of the transformer to follow the induced structure.

Finally, we apply document attention to produce the weighted sum of all the updated local focus representations. The document attention identifies relative weights of updated representations which enables our model to handle any document length.

4 Experiments

4.1 Implementation Details

We implement our model using the PyTorch library and use the Stanford Stanza library[2] for sentence tokenization. We employ XLNet for the pretrained language model. For the baselines which do not employ a pretrained language model (Dong et al., 2017; Mesgar and Strube, 2018), GloVe is employed for word embeddings, trained on Google News (Pennington et al., 2014) (see Appendix A for more details).

To compare baselines within the same framework, we re-implement all of them in PyTorch. We then use our re-implementation to report the performance of models with 10 runs with different random seeds. We verify statistical significance (p-value<0.01) with both a one-sample t-test, which verifies the reproducibility of the performance of each model, and a two-sample t-test, which verifies that the performance of our model is statistically significantly different from other models.

Within same framework we compare the size of models used in our experiment. Our neural model uses a number of parameters comparable to the state of the art, the transformer-based model (Moon et al. (2019): 118M < Jeon and Strube (2020): 136M < Our model: 137M).

[2]https://stanfordnlp.github.io/stanza

4.2 Baselines: Neural Coherence Models

In all three downstream tasks, we compare our model against recent neural coherence models. First, Mesgar and Strube (2018) propose a neural local coherence model, based on Centering theory. This model connects the most related states of a Recurrent Neural Network, then represents the coherence patterns using semantic distances between the states. Second, Moon et al. (2019) propose a unified neural coherence model to consider local and structural aspects. This model consists of two modules when they employ a pretrained language model (Peters et al., 2018): a module of inter-sentence relations using a bilinear layer and a topic structure module applying a depth-wise convolutional layer to the sentence representations. To ensure fair comparison, XLNet is employed for this model as well, instead of ELMo (Peters et al., 2018).

	Avg Accuracy
Moon et al. (2019)-1SentEnc	**91.40**
Our Model	86.45

Table 1: Shuffle Test: Mean (standard deviation) accuracy performance of shuffling test on GCDC, averaged on four domains. 1SentEnc indicates that each sentence is encoded separately on the pretrained language model.

4.3 Artificial Task: Shuffle Test

We first evaluate our model on the artificial setup, the shuffle test, used in the earlier works. We follow the setup used in Lai and Tetreault (2018). In this setup, our model outperforms a simple neural model relying on the pretrained language model. Moon et al. (2019) evaluate their models only in this setup. It achieves outstanding performance in this setup. However, in the following sections, our results show that this model does not outperform our model in downstream tasks.

Our results are not surprising. There is a line of recent work which shows that this setup is not desirable to evaluate coherence from diverse perspectives. Laban et al. (2021) show that employing fine-tuned language models simply achieves a near-perfect accuracy on this setup. O'Connor and Andreas (2021) measure usable information by selectively ablating lexical and structural information in transformer-based language models. Their findings show that prediction accuracy depends on information about local word co-occurrence, but

Model	Prompt								Avg Acc
	1	2	3	4	5	6	7	8	
Dong et al. (2017)	69.30	66.47	65.84	66.38	68.89	64.20	67.11	65.73	66.74
Mesgar and Strube (2018)	56.25	55.94	55.20	57.20	56.57	55.10	56.97	58.39	56.45
Averaged-XLNet-1SentEnc	70.73	69.48	68.98	67.52	72.35	70.94	70.14	69.01	69.89
Moon et al. (2019)-1SentEnc	73.75	72.13	72.92	73.29	75.12	74.69	72.89	72.09	73.36
Jeon and Strube (2020)-1SentEnc	75.10	73.35	74.75	74.18	76.38	74.30	73.61	73.44	74.39
Jeon and Strube (2020)-2SentsEnc	76.35	75.40	75.00	74.85	77.63	74.06	73.71	74.00	75.12
Our Model	**78.38**	**75.70**	**76.58**	**76.56**	**79.10**	**76.41**	**75.03**	**74.57**	**76.54**

Table 2: AES: TOEFL Accuracy performance comparison on the test sets, where 1SentEnc indicates that sentences are encoded individually and 2SentsEnc indicates that two adjacent sentences are encoded at once on the pretrained language model (see Appendix C for more details).

not word order or global position. We suspect that exploiting all information of a sentence is more beneficial to shuffle tests than entity-based modeling. Based on these findings, we evaluate our model on three downstream tasks used for evaluating coherence models, automated essay scoring, assessing writing quality, and assessing discourse coherence. We encourage future work not to evaluate coherence models on the artificial setup solely.

4.4 Automated Essay Scoring (AES)

Dataset. To evaluate the coherence models on AES, we evaluate them on the Test of English as a Foreign Language (TOEFL) dataset (Blanchard et al., 2013). While the Automated Student Assessment Prize (ASAP) dataset[3] is commonly used for AES, TOEFL has generally higher quality of essays compared to essays in ASAP. The prompts in ASAP are written by students in grade levels 7 to 10 of US middle schools. Many essays in ASAP consist of only a few sentences. In contrast, the prompts in TOEFL are submitted for the standard English test for the entrance to universities by non-native students. The prompts in TOEFL do not vary so much, the student population is more controlled, and essays have a similar length (see Appendix A for more details).

Evaluation Setup. We follow the evaluation setup of previous work on AES (Taghipour and Ng, 2016). For TOEFL, we evaluate performance with accuracy for the 3-class classification problem with 5-fold cross-validation. We use the same split for the cross-validation, used by Jeon and Strube (2020). The cross-entropy loss is deployed for training. The ADAM optimizer is used for our model with a learning rate of 0.003. We evaluate performance for 25 epochs on the validation set with a mini-batch size of 32. The model which reaches the best accuracy on the validation set is then applied to the test set.

Baselines. We compare against Dong et al. (2017), a neural model proposed for AES. They present a model which consists of a convolutional layer, followed by a recurrent layer, and an attention layer (Bahdanau et al., 2015) between the adjacent tokens.

Results. Table 2 reports the performance on TOEFL. Dong et al. (2017) report better performance than the more recent neural model based on Centering theory (Mesgar and Strube, 2018). A simple model relying on the pretrained language model outperforms this model, which averages all sentences to a vector representation (henceforth, Avg-XLNet). Moon et al. (2019) show that their unified model outperforms previous models on the artificial task, the shuffle test. However, it does not outperform the previous models on the AES task. Jeon and Strube (2020) outperform previous models. Finally, our model, which integrates local and structural aspects, achieves state-of-the-art performance. We perform an ablation study to investigate the contribution of individual components. We compare with Jeon and Strube (2020) who encode two adjacent sentences using the pretrained language model (2SentsEnc). Our results verify that this encoding improves performance, but our model benefits from the novel local coherence module even more.

4.5 Assessing Writing Quality (AWQ)

Dataset. Louis and Nenkova (2013) create a dataset of scientific articles from the New York Times (NYT) for assessing writing quality. They assign each article to one of two classes by a semi-supervised approach: typical or good. Though articles included in both classes are of good quality overall, Louis and Nenkova (2013) show that linguistic features contribute to distinguish different

[3] https://kaggle.com/c/asap-aes

Model	Yahoo	Clinton	Enron	Yelp	Avg Acc
∗Li and Jurafsky (2017)	53.5	61.0	54.4	49.1	51.7
Mesgar and Strube (2018)	47.3 (1.8)	57.7 (0.6)	50.6 (1.2)	54.6 (0.3)	52.6
∗Lai and Tetreault (2018)	54.9	60.2	53.2	54.4	55.7
Avg-XLNet-1Sent	**58.0** (3.9)	57.6 (0.3)	54.3 (0.8)	55.9 (0.4)	56.4
Moon et al. (2019)-1SentEnc	56.2 (0.5)	61.0 (0.4)	53.6 (0.5)	56.6 (0.4)	56.9
Jeon and Strube (2020)-1SentEnc	56.4 (0.6)	62.5 (0.9)	54.5 (0.4)	56.9 (0.3)	57.6
Jeon and Strube (2020)-2SentsEnc	57.2 (0.5)	63.0 (0.4)	54.4 (0.4)	**56.9** (0.2)	57.9
Our Model	58.4 (0.2)	**64.2** (0.4)	**55.3** (0.3)	57.3 (0.2)	**58.9**

Table 3: ADC: Mean (standard deviation) accuracy performance on the test sets in GCDC (∗: reported performance in Lai and Tetreault (2018)).

classes of writing quality.

Evaluation Setup. For NYT, we follow the setup used in previous work. Louis and Nenkova (2013) and Ferracane et al. (2019) undersample the dataset to mitigate the bias of the uneven label distribution. Following Ferracane et al. (2019), Jeon and Strube (2020) partition the dataset into 80% training, 10% validation, and 10% test set, respectively. We use the ADAM optimizer with a learning rate of 0.001 and a mini-batch size of 32. We evaluate performance for 25 epochs.

Baselines. Liu and Lapata (2018) propose a neural model which induces structural information without a labeled resource. It induces the non-projective dependency structure by structured attention.

	NYT
Liu and Lapata (2018)-reimplemented	54.35 (1.00)
Averaged-XLNet-1SentEnc	67.53 (3.48)
Moon et al. (2019)-1SentEnc	74.75 (1.27)
Jeon and Strube (2020)-1SentEnc	75.12 (1.10)
Jeon and Strube (2020)-2SentsEnc	76.43 (0.88)
Our Model	**77.52 (0.42)**

Table 4: AWQ: Mean (standard deviation) accuracy performance of assessing writing quality on the test sets in NYT.

Results. Table 4 shows the performance on NYT. Ferracane et al. (2019) reported the best performance of the latent learning model for discourse structure (Liu and Lapata, 2018) on NYT. However, Jeon and Strube (2020) show that the good results are due to the embeddings trained from the target dataset. They also report that Avg-XLNet outperforms this model which employs Glove embeddings. Moon et al. (2019) show better performance than this simple model, but it does not outperform Jeon and Strube (2020). Our model achieves state-of-the-art performance. An ablation study of the joint sentence encoding verifies that our model gains improvement not only from this encoding but also from our local coherence module.

4.6 Assessing Discourse Coherence (ADC)

Dataset. While previous work evaluates coherence models on formally written texts (Barzilay and Lapata, 2008), GCDC (Lai and Tetreault, 2018) is designed to evaluate coherence models on informal texts, such as emails or online reviews. The dataset contains four domains: Clinton and Enron for emails, Yahoo for questions and answers in an online forum, and Yelp for online reviews of businesses. The quality of the dataset is controlled to have evenly-distributed scores and a low correlation between discourse length and scores[4].

Evaluation Setup. For GCDC, we perform the experiments following previous work (Lai and Tetreault, 2018). We perform 10-fold cross-validation, use accuracy as evaluation measure on the 3-class classification, and use the cross-entropy loss function.

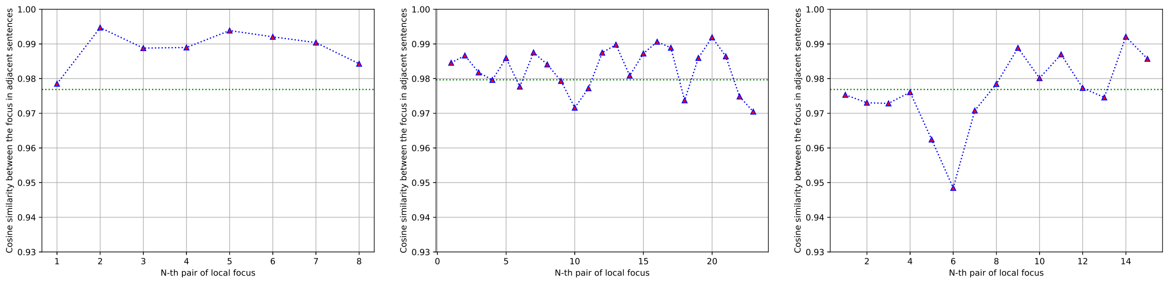
Baselines. Li and Jurafsky (2017) propose a neural model based on cliques, that are sets of adjacent sentences. This model uses the cliques taken from the original article as a positive label and uses cliques with randomly permutated ones as a negative label. Lai and Tetreault (2018) show that a simple neural model which uses paragraph information outperforms previous models on GCDC.

Results. Table 3 summarizes the performance on GCDC. While Avg-XLNet outperforms previous baselines, other advanced neural models show similar performance. Our model performs slightly better than Jeon and Strube (2020) with two sentences encoding. This shows that the gains mainly benefit from this encoding strategy. We suspect that Jeon and Strube (2020) do not benefit from structural information since texts on GCDC are not well-organized. The texts mostly consist of a few sentences, and they express the writers' emotion. Based on this, Lai and Tetreault (2018) state that

[4]The Pearson correlation between text length and scores is lower than 0.12 in all domains.

TOEFL-P1 (%)	TOEFL-P1-**NP** (%)	NYT-1516415 (%)	NYT-1516415-**NP** (%)
_broad (3.63)	i (5.45)	_theory (4.03)	it (4.96)
_many (1.79)	you (2.74)	_universe (3.22)	we (4.13)
_special (1.50)	broad knowledge (2.64)	_said	the universe (2.48)
i (1.47)	it (2.38)	stan (2.42)	he (2.48)
_specialize (1.46)	we (1.74)	ein (2.42)	physics (1.65)
_know (1.05)	knowledge (1.34)	dr (2.42)	space (1.65)
_specialized (0.99)	he (1.30)	_do (2.42)	string theory (1.65)
_knowledge (0.90)	people (1.20)	_can (1.61)	life (1.65)
_academic (0.90)	they (1.17)	_extra (1.61)	i (1.65)
_major (0.65)	many academic subjects (0.95)	_co (1.61)	dimensions (1.65)

Table 5: ADC: Comparison of focus on any items and noun phrases on top-10 most preferred centers (proportions) of essays submitted to prompt 1 in TOEFL and a NYT article ID 1516415 (see Appendix D for more details).



(a) TOEFL: P1, low score, id: 10226 (b) TOEFL: P1, high score, id: 63719 (c) TOEFL: P1, low score, id: 598381

Figure 2: Semantic consistency on TOEFL. The green horizontal line indicates the average of semantic similarities between local foci. The blue line indicates the semantic similarities between adjacent local foci. A semantic transition occurs when the semantic similarity between the local foci is lower than the green line. Texts of lower quality are mostly semantically too consistent (id:10226) or too variant (id:598381).

texts of lower quality have sudden topic changes. We also suspect that human annotators recognize important entities in the texts, such as the name of a person in the US government.

5 Analysis

5.1 Capturing Focus Using Entities

In Centering theory, the focus is described as the most important item in a sentence. Jeon and Strube (2020) capture the focus using attention weight scores and analyze texts assigned to different qualities using this focus. They state that the focus is difficult to interpret when it is composed of sub-words. To investigate this further, we compare the focus captured on (sub-)words and the focus constrained to entities. Table 5 indicates that constraining focus to entities leads to better explainability, in particular on NYT (see Table 11 in the Appendix D for more details). For example, in the NYT-1516415 news article about the String theory, a subword of "ein" is not interpretable focus while it may represent a useful representation in the vector space for a neural model. In contrast, our entity-based modeling leads our model to better explainability. Instead of "ein", it provides more interpretable focus, "Einstein", a theoretical physicist. In TOEFL, "many academic subjects" is more interpretable focus than focus consists of a single subword token either "many" or "subjects". Table 5 also shows that our model mainly uses pronouns, and noun phrases are playing an important role in the text to represent focus. Our findings suggest that further investigation is needed to understand how pretrained language models work on pronouns to process a long text.

5.2 Local Coherence Patterns

Using interpretable focus information, we investigate differences in focus transitions of texts assigned to different scores. Motivated by the definition of the continue and the shift transition in Centering theory, we define semantic consistency which represents the degree of semantic changes between local foci. Two adjacent sentences are semantically consistent when the semantic similarity (sim_i) between the local foci (lf) is higher than a semantic threshold ($\theta_{sem;score}$). This threshold is determined as the average of semantic similarities between local foci of adjacent sentences in the texts assigned the same score. Otherwise, a semantic transition (st) occurs between the local foci: $st_i = 1$ if $sim_i < \theta_{sem;score}$. Finally the semantic consistency (SC) is defined as follows: $SC = 1 - (count(st_i)/|lf|)$.

Figure 2 illustrates the semantic consistency on TOEFL, and Table 6 shows the statistics of the semantic consistency on texts assigned to different scores. Texts assigned a high score show lower semantic consistency on average. This indicates that texts of higher quality are overall more semantically variant than texts of lower quality. Additionally, we observe that texts assigned a low score show significantly larger proportions of an extreme level of semantic consistency. We define the extreme level as either texts whose semantic consistency is lower than 5%, indicating texts are highly variant, or texts whose semantic consistency is higher than 75%, indicating texts are highly consistent. Hence, these findings indicate that texts of lower quality are semantically too variant or too consistent. Texts of higher quality are are neither too variant nor too consistent.

We next inspect the focus of texts assigned to different scores (see Table 12,13, and 14 in the Appendix D for more details). It shows that the proportion of pronouns is higher in the local focus compared with their proportion in the focus captured on a sentence solely. The essays in TOEFL are argumentative essays, and good essays should use facts and evidence to support their claim (Wingate, 2012). We observe that texts assigned a low score frequently include claims without convincing evidence. This causes our model to capture focus on pronouns more frequently in these texts. In contrast, texts assigned a high score include convincing evidence to support claims, and this lets our model capture different types of foci in these texts.

5.3 Error Analysis

Finally, we conduct an error analysis to investigate how our model works differently compared to previous coherence models on TOEFL. We first compare the predicted scores with Moon et al. (2019) and a simple model which only considers context, averaged-XLNet. These two baselines show biased predictions on the middle score. We suspect that this is caused by the label bias in TOEFL (Blanchard et al., 2013). Biased label distributions cause biased predictions, and they benefit from these biased predictions. In contrast, our model benefits more from predicting high scores correctly as well as other scores, indicating our coherence model assess text quality better.

	S_{Low}	S_{Mid}	S_{High}
Avg SC (std)	55.87 (24.53)	54.45 (21.38)	54.05 (19.70)
Prop of Ext level	17.63	11.54	8.59

Table 6: Semantic consistency statistics (%) for the texts assigned to different scores (S). An extreme level (Ext) is defined as either semantic consistency to be lower than 5% (semantically too variant) or higher than 75% (semantically too consistent).

We then compare with the previous state of the art (Jeon and Strube, 2020). This baseline induces discourse structure to model structural coherence. It captures semantic relations between discourse segments, not just between adjacent sentences. We observe two error cases when this baseline struggles to predict correctly. It predicts scores lower than the ground-truth score for texts which lack support and evidence for claims. However, these texts have a well-organized paragraph for one or two claims. We suspect that this leads human annotators to assign a mid or a high score though the text is not well-organized overall. In contrast, it predicts scores higher than ground-truth scores when unrelated claims are listed or claims are listed without evidence. Our model, which captures local coherence between adjacent sentences, deals with these cases better (see Table 15 and 16 in the Appendix D for more details).

6 Conclusions

We propose a neural coherence model based on entities by constraining the input to noun phrases. It leads our model to better explainability and to set a new state of the art in end applications. It also allows us to reveal that texts of higher qualities are neither semantically too consistent nor too variant.

Our findings suggest a few interesting directions for future work. As our model sets a new state of the art by constraining models to focus on entities, we could design more efficient modeling instead of considering all information on other tasks as well. Our analysis shows that pretrained language models frequently exploit coreference relations to capture semantic relations. We could design an advanced neural model which exploits these relations explicitly, which could lead to better explainability and better understanding of how transformer-based models work.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *In Proceedings of the ICLR Conference*.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1251–1258.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162, Vancouver, Canada. Association for Computational Linguistics.

Vanessa Wei Feng and Graeme Hirst. 2012. Extending the entity-based coherence model with multiple ranks. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 315–324, Avignon, France. Association for Computational Linguistics.

Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Katrin Erk. 2019. Evaluating discourse in structured text representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 646–653, Florence, Italy. Association for Computational Linguistics.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Camille Guinaudeau and Michael Strube. 2013. Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 93–103, Sofia, Bulgaria. Association for Computational Linguistics.

Daniel Hardt and Maribel Romero. 2004. Ellipsis and the structure of discourse. *Journal of Semantics*, 21(4):375–414.

Sungho Jeon and Michael Strube. 2020. Centering-based neural coherence modeling with hierarchical discourse segments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7458–7472, Online. Association for Computational Linguistics.

Aravind K Joshi and Scott Weinstein. 1981. Control of inference: Role of some aspects of discourse structure-centering. In *IJCAI*, pages 385–387.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Philippe Laban, Luke Dai, Lucas Bandarkar, and Marti A Hearst. 2021. Can transformer models measure coherence in text? Re-thinking the shuffle test. *arXiv preprint arXiv:2107.03448*.

Alice Lai and Joel Tetreault. 2018. Discourse coherence in the wild: A dataset, evaluation and methods. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 214–223, Melbourne, Australia. Association for Computational Linguistics.

Jiwei Li and Dan Jurafsky. 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.

Yang Liu and Mirella Lapata. 2018. Learning structured text representations. *Transactions of the Association for Computational Linguistics*, 6:63–75.

Annie Louis and Ani Nenkova. 2013. What makes writing great? first experiments on article quality prediction in the science journalism domain. *Transactions of the Association for Computational Linguistics*, 1:341–352.

Mohsen Mesgar and Michael Strube. 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.

Tasnim Mohiuddin, Prathyusha Jwalapuram, Xiang Lin, and Shafiq Joty. 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.

Han Cheol Moon, Tasnim Mohiuddin, Shafiq Joty, and Chi Xu. 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.

Joe O'Connor and Jacob Andreas. 2021. What context features can transformer language models use? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 851–864, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Leila Pishdad, Federico Fancellu, Ran Zhang, and Afsaneh Fazly. 2020. How coherent are neural models of coherence? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6126–6138, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Candace Sidner. 1983. Focusing in the comprehension of definite anaphora. *Computational models of discourse*, pages 267–330.

Candace L. Sidner. 1981. Focusing for interpretation of pronouns. *American Journal of Computational Linguistics*, 7(4):217–231.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. *In Proceedings of the ICLR Conference*.

Dat Tien Nguyen and Shafiq Joty. 2017. A neural local coherence model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1320–1330, Vancouver, Canada. Association for Computational Linguistics.

Yaushian Wang, Hung-Yi Lee, and Yun-Nung Chen. 2019. Tree transformer: Integrating tree structures into self-attention. 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 1061–1070, Hong Kong, China. Association for Computational Linguistics.

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020. Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235, Online. Association for Computational Linguistics.

Ursula Wingate. 2012. 'Argument!'Helping students understand what essay writing is about. *Journal of English for Academic Purposes*, 11(2):145–154.

Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. In *International Conference on Learning Representations*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Lang Yu and Allyson Ettinger. 2020. Assessing phrasal representation and composition in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4896–4907, Online. Association for Computational Linguistics.

A Training and Parameters

For the three datasets, we use a mini-batch size of 32 with random-shuffle. The ADAM optimizer is used to train our models with a learning rate of 0.001 and epsilon of 1e-4. We evaluate performance for 25 epochs. For the baseline models which do not use a pretrained language model, we use Glove pretrained embeddings with 100-dimensional for TOEFL and with 50-dimensional for NYT. We clip gradients by 1.0. To update sentence representations obtained by a pretrained language model, we use the same dimension of the pretrained language model on a tree-transformer. We manually tune hyperparameters.

We encode adjacent two sentences at once using XLNet instead of the whole document at once. Our dataset consists of long documents i.e., journal articles with more than 3,000 tokens. For employing the pretrained model, it is practically infeasible to encode all words in a document at once due to memory limitations. We use 23GB GPU memory a NVidia P40 on ADC and AES and 46GB GPU memory of two NVidia P40s for each run on AWQ. For training our model, it takes approximately 0.8 days on TOEFL, 6.5 days on NYT, and 0.6 days on GCDC.

B Data Description Details

Table 7 describes statistics on two datasets, TOEFL[5] and NYT[6]. We split a text at the sentence level by Stanford Stanza library, and tokenize them by the XLNet tokenizer. Table 8 describes the topic of each prompt in TOEFL. They are all open-ended tasks, that do not have given context but require students to submit their opinion.

C Evaluation: Shuffle Test

The shuffle test is introduced to evaluate coherence models with other tests in the previous work (Barzilay and Lapata, 2008).

Barzilay and Lapata (2008) introduce the shuffle test to evaluate coherence models with other tests.

D Evaluations Details

E Analysis Details

Dataset	#Texts	Avg len (Std)	Max # tokens	Scores
G-Y	1,200	173 (48)	378	1-3
G-C	1,200	200 (65)	385	1-3
G-E	1,200	203 (67)	388	1-3
G-P	1,200	198 (58)	374	1-3
T-P1	1,656	401 (97)	902	1-3
T-P2	1,562	423 (97)	902	1-3
T-P3	1,396	407 (102)	837	1-3
T-P4	1,509	405 (99)	852	1-3
T-P5	1,648	424 (101)	993	1-3
T-P6	960	425 (101)	925	1-3
T-P7	1,686	396 (87)	755	1-3
T-P8	1,683	407 (92)	795	1-3
NYT	8,512	1,841 (1,221)	18,728	1-2

Table 7: Three Datasets statistics on tokenization: i) four domains in GCDC, Yahoo (G-Y), Clinton (G-C), Enron (G-E), Yelp (G-P), ii) each TOEFL prompt (T-P), and iii) NYT.

Prompt 1	Agree or Disagree: It is better to have broad knowledge of many academic subjects than to specialize in one specific subject.
Prompt 2	Agree or Disagree: Young people enjoy life more than older people do.
Prompt 3	Agree or Disagree: Young people nowadays do not give enough time to helping their communities.
Prompt 4	Agree or Disagree: Most advertisements make products seem much better than they really are.
Prompt 5	Agree or Disagree: In twenty years, there will be fewer cars in use than there are today.
Prompt 6	Agree or Disagree: The best way to travel is in a group led by a tour guide.
Prompt 7	Agree or Disagree: It is more important for students to understand ideas and concepts than it is for them to learn facts.
Prompt 8	Agree or Disagree: Successful people try new things and take risks rather than only doing what they already know how to do well.

Table 8: Topic description: TOEFL.

[5]https://catalog.ldc.upenn.edu/LDC2014T06
[6]https://catalog.ldc.upenn.edu/LDC2008T19

Model	Prompt 1	2	3	4	5	6	7	8	Avg Acc
Dong et al. (2017)	69.30 (0.41)	66.47 (0.58)	65.84 (0.56)	66.38 (0.56)	68.89 (0.38)	64.20 (0.64)	67.11 (0.59)	65.73 (0.31)	66.74
Mesgar and Strube (2018)	56.25 (0.72)	55.94 (0.44)	55.20 (0.75)	57.20 (0.16)	56.57 (0.49)	55.10 (0.39)	56.97 (0.56)	58.39 (0.29)	56.45
Averaged-XLNet-1SentEnc	70.73 (0.73)	69.48 (0.53)	68.98 (1.12)	67.52 (0.51)	72.35 (0.46)	70.94 (0.82)	70.14 (0.42)	69.01 (0.56)	69.89
Moon et al. (2019)-1SentEnc	73.75 (0.67)	72.13 (0.58)	72.92 (0.54)	73.29 (0.35)	75.12 (0.50)	74.69 (0.57)	72.89 (0.35)	72.09 (0.35)	73.36
Jeon and Strube (2020)-1SentEnc	75.10 (0.74)	73.35 (0.92)	74.75 (0.61)	74.18 (1.07)	76.38 (0.91)	74.30 (1.13)	73.61 (0.72)	73.44 (1.15)	74.39
Jeon and Strube (2020)-2SentsEnc	76.35 (0.44)	75.40 (0.75)	75.00 (0.34)	74.85 (0.50)	77.63 (0.40)	74.06 (0.37)	73.71 (0.25)	74.00 (0.63)	75.12
Our Model	**78.38** (0.42)	**75.70** (0.60)	**76.58** (0.46)	**76.56** (0.37)	**79.10** (0.35)	**76.41** (0.20)	**75.03** (0.32)	**74.57** (0.38)	**76.54**
Our Model+Coref	75.70 (0.60)	75.36 (0.63)	75.04 (0.37)	74.92 (0.60)	76.97 (0.51)	74.43 (0.72)	73.53 (0.69)	72.81 (0.38)	74.84

Table 9: TOEFL Accuracy performance comparison on the test sets (std), where 1SentEnc indicates that sentences are encoded individually and 2SentsEnc indicates that adjacent sentences are encoded at once on the pretrained language model.

Model	Prompt 1	2	3	4	5	6	7	8	Avg Acc
Averaged-XLNet-1SentEnc	71.06 (0.43)	70.56 (0.50)	67.17 (0.99)	67.02 (0.98)	71.42 (0.31)	69.76 (0.77)	68.54 (0.73)	68.72 (0.51)	69.28
Moon et al. (2019)-1SentEnc	74.31 (0.67)	71.15 (0.12)	72.83 (0.96)	73.71 (0.80)	74.94 (0.53)	73.89 (1.00)	72.18 (0.76)	72.04 (0.73)	73.13
Jeon and Strube (2020)-1SentEnc	73.76 (0.74)	71.09 (0.92)	72.57 (0.61)	71.86 (1.07)	73.87 (0.91)	71.08 (1.13)	71.49 (0.72)	71.46 (1.15)	72.15
Jeon and Strube (2020)-2SentsEnc	76.66 (0.50)	75.48 (0.68)	74.46 (0.74)	74.72 (0.36)	76.24 (0.50)	75.26 (0.53)	73.82 (0.43)	73.19 (0.67)	74.98
Our Model	**77.44** (0.59)	**75.48** (0.74)	**76.72** (0.72)	**76.57** (0.46)	**79.22** (0.61)	**75.89** (0.85)	**75.66** (0.77)	**74.33** (0.74)	**76.41**

Table 10: TOEFL Accuracy performance comparison on the validation sets (std), where 1SentEnc indicates that sentences are encoded individually and 2SentsEnc indicates that adjacent sentences are encoded at once on the pretrained language model.

TOEFL-P1-**NP** (%)	TOEFL-P2-**NP** (%)	TOEFL-P3-**NP** (%)	TOEFL-P4-**NP** (%)
i (5.45)	young people (5.57)	young people (5.26)	i (4.67)
you (2.74)	they (5.21)	i (4.71)	it (3.83)
broad knowledge (5.64)	i (4.42)	they (3.70)	they (3.61)
it (2.38)	life (4.12)	time (1.64)	advertisements (2.03)
we (1.74)	older people (2.70)	enough time (1.52)	products (1.96)
knowledge (1.34)	it (1.50)	it (1.46)	you (1.82)
he (1.30)	you (1.40)	their communities (1.23)	we (1.59)
people (1.20)	we (1.05)	people (1.19)	people (1.49)
they (1.17)	old people (1.02)	we (1.10)	most advertisements (1.10)
many academic subjects (0.95)	people (0.95)	them (0.92)	the product (0.96)
TOEFL-P5-**NP** (%)	TOEFL-P6-**NP** (%)	TOEFL-P7-**NP** (%)	TOEFL-P8-**NP** (%)
cars (4.54)	i (7.73)	i (5.16)	i (4.90)
i (4.25)	you (4.16)	ideas and concepts (3.74)	they (3.51)
twenty years (3.26)	a group (3.96)	facts (3.73)	you (2.70)
people (2.07)	a tour guide (3.49)	students (3.05)	he (2.24)
it (1.81)	we (2.36)	it (2.82)	it (2.22)
we (1.71)	it (2.20)	they (2.61)	successful people (2.13)
they (1.50)	they (1.45)	you (1.89)	people (2.01)
use (1.49)	people (1.39)	we (1.87)	risks (1.85)
today (1.13)	the best way (0.92)	them (1.10)	new things (1.76)
a car (0.75)	the tour guide (0.85)	the facts (1.09)	success (1.57)
NYT-1458761-**NP** (%)	NYT-1516415-**NP** (%)	NYT-1705265-**NP** (%)	NYT-1254567-**NP** (%)
i (3.82)	it (4.96)	i (4.79)	he (4.22)
colorado (3.82)	we (4.13)	he (4.79)	it (3.52)
2001 (2.29)	the universe (2.48)	they (3.42)	einstein (3.52)
montana (2.29)	he (2.48)	diet (2.74)	schrodinger's (2.82)
colorado springs 2004 (1.53)	physics (1.65)	cancer (2.74)	they (2.11)
denver (1.53)	space (1.65)	it (2.05)	itself (2.11)
qwest (1.53)	string theory (1.65)	breast cancer (2.05)	bohr (2.11)
we (1.53)	life (1.65)	people (2.05)	a physicist (1.41)
the state (1.53)	i (1.65)	those (2.05)	berlin (1.41)
jobs (1.53)	dimensions (1.65)	prostate cancer (1.37)	light (1.41)

Table 11: Top-10 most frequent focus (proportions) of essays submitted to the same prompt in TOEFL (see Appendix. A for given topics) and four articles in NYT whose id is 1458761, 1516415, 1705265, and 1254567, respectively. The title of NYT articles are as follows, 1458761: "Among 4 States, a Great Divide in Fortunes", 1516415: "One Cosmic Question, Too Many Answers", 1705265: "Which of These Foods Will Stop Cancer?", and 1254567: "Quantum Theory Tugged, And All of Physics Unraveled".

T-P1-Local-Low (%)	T-P1-Single-Low (%)	T-P1-Local-High (%)	T-P1-Single-High (%)
i (8.77)	i (6.44)	i (5.98)	i (5.05)
you (3.51)	broad knowledge (3.43)	you (3.23)	you (2.29)
it (3.42)	we (2.19)	it (2.70)	it (2.21)
one specific subject (2.58)	you (2.19)	one specific subject (1.73)	broad knowledge (1.84)
we (2.48)	it (2.13)	we (1.37)	we (1.65)
broad knowledge (1.78)	many academic subjects (1.42)	a broad knowledge (1.27)	knowledge (1.56)
many academic subjects (1.67)	he (1.42)	one (1.22)	he (1.22)
he (1.19)	they (1.24)	he (1.20)	they (1.11)
they (1.04)	knowledge (1.05)	this (1.17)	a broad knowledge (1.09)
that (0.08)	that (0.95)	many academic subject (1.16)	specialization (1.09)
T-P3-Local-Low (%)	T-P3-Single-Low (%)	T-P3-Local-High (%)	T-P3-Single-High (%)
i (8.97)	i (5.57)	young people (6.33)	young people (4.79)
young people (6.65)	young people (4.77)	i (5.91)	i (4.48)
they (5.53)	they (4.63)	they (4.35)	they (3.42)
the young people (2.72)	it (1.94)	it (1.98)	time (1.69)
it (2.44)	their communities (1.79)	the young people (1.91)	it (1.43)
enough time (1.96)	time (1.79)	the community (1.74)	enough time (1.24)
them (1.80)	enough time (1.65)	their communities (1.70)	their communities (1.18)
their communities (1.76)	we (1.18)	this (1.60)	people (1.18)
we (1.64)	them (1.13)	them (1.50)	we (1.05)
there (1.24)	the young people (1.04)	people (1.36)	them (0.89)
T-P7-Local-Low (%)	T-P7-Single-Low (%)	T-P7-Local-High (%)	T-P7-Single-High (%)
i (9.08)	i (5.95)	i (6.81)	i (5.29)
it (4.11)	ideas and concepts (3.70)	it (3.78)	ideas and concepts (4.16)
they (3.29)	facts (3.56)	facts (3.48)	facts (3.86)
we (3.09)	students (3.23)	ideas and concepts (3.23)	students (2.97)
facts (2.90)	they (3.14)	you (2.59)	it (2.90)
ideas and concepts (2.57)	it (1.95)	they (2.08)	they (2.36)
you (2.23)	we (2.34)	the facts (2.05)	you (2.13)
students (2.15)	ideas (1.69)	students (1.91)	we (1.60)
the students (1.68)	you (1.45)	a student (1.58)	them (1.25)
the facts (1.41)	them (1.26)	we (1.45)	ideas (1.06)
T-P8-Local-Low (%)	T-P8-Single-Low (%)	T-P8-Local-High (%)	T-P8-Single-High (%)
i (8.07)	i (5.45)	i (9.90)	i (4.56)
they (4.83)	they (4.73)	you (6.55)	they (2.88)
new things (3.91)	he (3.10)	they (5.16)	you (2.64)
you (2.75)	successful people (2.85)	new things (2.65)	it (2.09)
it (2.64)	new things (2.43)	it (2.30)	he (2.02)
he (2.64)	people (2.01)	he (1.90)	risks (1.94)
successful people (1.80)	you (1.88)	people (1.52)	success (1.78)
people (2.04)	it (1.59)	risks (1.49)	successful people (1.77)
we (1.45)	success (1.55)	successful people (1.44)	people (1.64)
success (0.74)	we (1.26)	we (1.44)	new things (1.47)

Table 12: Comparison of the top-10 the most frequent local focus, captured on the two adjacent sentences, (proportions) and single focus, captured on a sentence solely, of essays submitted to each prompt in TOEFL for the low and the high score (see Appendix. B for given topics).

#	Example text of low quality
1	**I**[1] absolutely agree about the many academic subjects are beneficial for knowledge, because it provide **lots of opportunities**[1,2], I mean it's good for our future.
2	In my experience, when **I**[3] was second grade in middle school, a teacher gave **a homework**[2] to us which was to find our talant.
3	**I**[3,4] tried to think what am I good at and what do I like.
4	However, **I**[4] couldn't, because I couldn't find **my talants**[5].
5	after my highschool finally, I found **my talants**[5].
6	**My talant**[6] is to study a law.
7	When **I**[6] was first grade in the highschool, **I**[7] had a friend who called Che-Jea-Heong.
8	He was **very special friend**[7,8].
9	He always tried to think **strange way**[8,9].
10	At first, **I**[9] didn't want to talk with him, but when **we**[10] talked about the talant, we became a friend.
11	Actually, **his father**[10,11] is police.
12	And **his family**[11] is very poor.
13	So, first **we**[12] started to talk his father.
14	why **he**[12,13] is poor.
15	After that **we**[13,14] began to think law.
16	Then **we**[14] found **our talant**[15].
17	Actually, **this**[16] I found **this talant**[15] from the school project.
18	When **I**[16,17] was 3grade in middle school, I took a class which was Korean language class, in the class, we had a special study which was law.
19	Because, **my teacher**[17,18] thought law is beneficial for stundent.
20	So **we**[18] tried to study **the law**[19] just one semester with a game.
21	However, my friends are really bored about this, but **me**[20] I really enjoyed **that law class**[19].
22	So after that semester, **I**[20,21] asked the teacher to study more laws, but she couldn't, because lots of people didn't like that.
23	Anyway, **I**[21,22] really like the law, also I'll study law in the university.
24	From this semester, **I**[22,23] can think many way to find my talant from the school subjects.
25	**I**[23,24] can think math, science, music or art.
26	So **we**[24,25] can have our opportunities.
27	Now days, many students cannot understand the school about the acadmic subjects that why they have to learn **too much subject**[25,26].
28	**I**[26] was too, but now I understand the school. And I really thanks from the school.

Table 13: Local focus on an example text assigned to the low score. The example is rewritten by us following the texts in TOEFL due to the non-public license. Bold style indicates local focus identified in our sentence encoding strategy, which encodes adjacent sentences at once. Superscripts indicate the order of this encoding.

#	Example text of high quality
1	Getting **more knowledge**[1,2] could expand ones boundary; serve as a parth to discover ones true passion; allow us to talk to other people and be capable of understanding the world around us.
2	Firstly, getting **more knowing**[2] of many academic subject areas could expand our boundaries because we know different subjects in **different fields**[3].
3	Each subject has its own uniquness, therefore **it**[4] would be beneficial to know a bit about **each areas**[3].
4	Secondly, exploring **more knowledge**[4,5] could serve as a path for people to discover their true passion.
5	Sometimes if we stay 'inside the box', it would be difficult for us to find other ways and have **the oppurtunity**[5,6] to think whether it was truly their passion or not.
6	When **I**[6] was in Grade 11, **I**[7] took courses in different areas, such as Chemistry, Accounting, Physical Education, Business, History etc.
7	**I**[7] wans't sure of what I wanted to study in university, and I don't want to limit my area of study, therefore **I**[8] decided to broaden my knowledg by taking many acdemic subjects.
8	However my friend, who seriously wanted to become a doctor, took **all science courses**[8,9], because she wanted to explore her passion.
9	As a result, I believe it would be better to have a broad knowledge of **many subjects**[9,10] before specializing one, unless you have found something that you really want to pursue.
10	Moreover, by studying **more subjects**[10], **it**[11] makes people easy to dive in conversations with new people.
11	Everyone have different backgrounds, therefore if you have knowledge from **different areas**[12], **it**[11] could be easier to socialize with people whom have different fields from we have.
12	A way of knowing **more subjects**[12,13] can be to read every section of the newspaper such as Businss, World, Entertainment etc.
13	This could help us to know **more knowledge**[13] and therefore we can be more talkative meeting **new people**[14].
14	Since **the world**[15] changes everyday, everyday **something new** [14]will happen.
15	If we don't have the basic background of **a certain subject**[15,16], we cannot understand others.
16	Moreover, a lot of subjects are tied on each other, therefore you will need knowledge from **other areas**[16,17] to understand the material better.
17	For example, business ties with politics, political changes could affect the business environment, henceforth it is mandatory for us to have **a simple background**[17,18] of politics to understand the changes of business around the world.
18	In conclusion, with all the reasons discussed so far, I believe that it is better to have broad knowledge of **many acadmic subjects**[18] than specializing in one specific subjects.

Table 14: Local focus on an example text of high quality. The examples is rewritten by us following the texts in TOEFL due to a non-public license. Bold style indicates local focus identified in the sentence encoding, which encodes two sentences at once. Superscripts indicate the order of this encoding.

Error Type	Example Essay
C_1	In my opinion is better to have a knowledge specialize in one particular subject since this is better to know a thing as well as you can. This is true in all the experiences of the life: refered to the university, e.g., the italian university, we can take the example of the of the two years of specialization. An other example we can see in a top-tier company, in fact each people that there are in this have a specific work to do and this bring to an excellent final operation. A person that are magnifically prepared on one thing will arrive at a sicure result because that ""is your bred""; we can also observe that the most good professors, scientists, sport players are all specialize on that they work and do not specialize on many works. We can also observe that the colloboration of great brains, each of them specialized on a thing, is important in many ways of the our life.
C_2	I strongly agree with the statement that knowing several subjects and being polyvalent in various fields is much more important that specializing in one area. These days, things are changing so fast that the moment you start a career or a specialization, the minute the facts and figures of the subject have changed. This essence of broad knowledge is what makes people succeed in the world. Unless you are 100% sure that you vocationally desire to specialize in a subject, the risk of not finding a suitable job because of the deviation of job offering is too high. Both with respect to time and money. For example, imagine that you decide to study IT sometime around the Internet boom. After you finish the 5 years of studying, you get out to society with high hopes and great expectations and suddenly you realize that the world does not need for IT people anymore because the market crashed down! Then you would most probably regret not to have chosen a more general Engineering degree such as an Electronical Engineering degree. Take the example of a devoted music students that really loves to play music to the point that they drop classes so they can go and play their music. Perhaps, they will become a succesful singer or solo player, but the chances that they fail are there and when that really comes true, they will not be able to attend university classes because they didn't passed high-school. Good and innovative ideas often are the result of composing other ideas. If on one side, you know how pollution of carbon dioxide is chemically produced and on the other, you are an expert on plant species, perhaps you can find a way to create a system to purify the air in the world. And moreover, if you have skills of marchandising and marketing, you can probably be in the Forbes' next month main page. Think that you can always specialize in the future. Going from the trunck of a tree to the tip of a branch is easy, but getting from one tip to another tip is, literally, as going back in time.

Table 15: Example Essays for Error Cases (C_1, C_2) on TOEFL (the examples are rewritten by us following the texts in TOEFL due to the non-public license). For texts corresponding to the C_1, Jeon and Strube (2020) predicts a low score and our model predicts a mid score ($C_1 : S_{JS} = L, S_O = M$). For texts corresponding to the C_2, Jeon and Strube (2020) predicts a mid score and our model predicts a high score ($C_2 : S_{JS} = M, S_O = H$).

Error Type	Example Essay
C_3	It seems difficult to choose one direction, becuse they are also have colorful life between the young people and the older people, but it does not mean, they are similar to me. I would like to agree with the young people enjoy life more than older people do, if a personal quality can be considerated as criterion to choose things. First of all, nowadays, era of information, many young people enjoy their life via the internet, even everything is possible in the digital industy. For instance, if a grandson of the older people live abroad, and the communication between the grandson and the grandfather is only via the telephone instead the internet online chatting what is cheaper than the international telephone call, but the older people can not use the internet, even they can not use a computer. On the other hand, the young people can adapt an environment quickly, so that they can migrate to another city for the different experience. most of older persons can not accept the different enviroment and what they will eat in the different areas, if the older person migrate to other citys or countries, they will be illness easier. The important things determining the young people enjoy life better is that they are educated in the signifcant era of information, so they are developed with the world development. For all mentioned above is why I agree with the statement that young people enjoy life more than older people do. Now, I do strongly agree with the statement.
C_4	Yes, it is better to have a broad knowledge of many academic subjects than specialisze in one specific area because of various reasons. If people have knowledge about a particular subject,it is good. But if they want to refrain themselves from foraying from other subjects they should make sure that they are very thorough with that subject.Because finally they should find a job on that basis only and more ver all the academic topics are interconnected so, it imperative to have knowledgein various fields. The above option would be good only if they find a job. They should always keep in mind the different possibilities in their carreer. They should ask themselves ""what if i dont get a job in my desired field of study?"" For instance I am a mechanical engineering student. as every one knows there is a difficult of getting jobs for mechanical engineers.if i continue with the same field would be left unemployed.Here I need to have an alternate option.I have my alternate option as computer sciences .I started learning some computer subjects.Now even if i do not get a job in my field of study, i may have a chance of getting it in field of computers.This would not leave me unemployed.I personally feel that being employed is better than being unemployed. This criteria not only works for two fields of same backround, it also works for a technical background and an arts background. For example, an electrical engineer who does not have a job and whose hobby is singing , can survive by giving some stage shows . Which would also be considered as an employment. Additionally, broader knowlege would not leave you speechless when you are in a group. Because when a group is discussing a topic and if you are silent , you may feel embarassing with that. But if you are familiar with the topic you can also give your opinion on the topic. this is possible only if you do not confine yourself to a particular field. Therefore, I conclude that having a broad knowledge is better than to specialize in one subject.

Table 16: Example Essays for Error Cases (C_3, C_4) on TOEFL (the examples are rewritten by us following the texts in TOEFL due to the non-public license). For texts corresponding to the C_3, Jeon and Strube (2020) predicts a mid score and our model predicts a low score ($C_3 : S_{JS} = M, S_O = L$). For texts corresponding to the C_4, Jeon and Strube (2020) predicts a high score and our model predicts a mid score ($C_4 : S_{JS} = H, S_O = M$).