Towards an Implementation of Rhetorical Structure Theory in Discourse Coherence Modelling

Michael Lambropoulos
School of Computer Science
The University of Sydney
mlam3772@uni.sydney.edu.au

Shunichi Ishihara
Speech and Language Laboratory
Australian National University
shunichi.ishihara@anu.edu.au

Abstract

In this paper, we combine the discourse coherence principles of Elementary Discourse Unit segmentation and Rhetorical Structure Theory parsing to construct meaningful graph-based text representations. We then evaluate a Graph Convolutional Network and a Graph Attention Network on these representations. Our results establish a new benchmark in F1-score assessment for discourse coherence modelling while also showing that Graph Convolutional Network models are generally more computationally efficient and provide superior accuracy.

1 Introduction

Natural Language Processing (NLP) has seen significant advancements, particularly with attention-based transformer models excelling in tasks such as machine translation, language modelling (Devlin et al., 2018), and sentiment analysis (Yang et al., 2019). However, effectively modelling discourse coherence remains a challenge, especially as long context and long form text genearation tasks become more prevalent. This research aims to address this by extending a graph-construction approach developed by Liu et al. (2023), integrating the linguistically-focused principles of Elementary Discourse Unit (EDU) segmentation and Rhetorical Structure Theory (RST) parsing into a graph-based approach using Graph Convolutional Network (GCN) and Graph Attention Network (GAT) architectures. This graph-based approach marks a departure from typical discourse coherence assessments such as those by Moon et al. (2019) which treat coherence as a sentence-rearrangement task. Our goal is to further the field of discourse coherence modelling, which is crucial for tasks like essay grading, mental health detection, and identifying machine-written text.

1.1 Motivation

Following recent breakthroughs in NLP, scientific research has focused on creating human-understandable output for text generation and classification tasks. The motivations behind such research are twofold. Firstly, human-computer interaction is predicated on two-way communication, meaning that whatever makes language understandable or believable is a standard of achievement to be attained. Secondly, Large Language Models (LLMs) are being seen as the embodiment of the language function of human processing capabilities. It then becomes a priority to imbue these models with human-like reasoning capabilities. As such, we seek to investigate to what extent the "coherence" of a piece of text can be adequately represented and assessed. Outlined by Jurafsky and Martin (2000), discourse coherence refers to the intelligibility of a text based on a range of factors including its structural arrangement and persistence of relevant topics throughout its paragraphs, sentences

1.2 The Need for Coherence in Generated Text

At the fringe of these discoveries is an area that requires both the technical oversight of NLP skills and an intimate knowledge of how meaning is conveyed in utterances (Ishibashi et al., 2023). It has been noted in current research (Wei et al., 2022; Wang et al., 2022) that language models still lack some fundamental process that can make freely generated text output unique, non-repetitive, relatively unpredictable, and relevant to the topic matter.

1.3 Research Aims

We observe in the literature that two core principles of coherence – local (paragraph level) and global coherence (structural composition) – are al-

most never combined in analysis. Many of the current state-of-the-art models seemingly disregard linguistic theory in favor of similarity and vector-based representations of discourse components, i.e., words and sentences, such as recent neural coherence work (Wang et al., 2017; Xu et al., 2019; Moon et al., 2019), with only recent work by that of Jiang et al. (2021) which aims for interesting synthesis of a sentence-embedding approach and a dimension grid Barzilay and Lapata (2008) model. Our study aims to address this gap by combining the linguistic principles of Elementary Discourse Unit (EDU) segmentation and Rhetorical Structure Theory (RST) parsing (discussed further in Sections 2.1 & 2.2) to construct more meaningful, graph-based representations of text for coherence modeling.

The main aims of this research are:

1. To evaluate whether the incorporation of linguistic theory principles (EDU segmentation and RST parsing) improves the performance of coherence modeling tasks.

2. To establish a significant improvement in performance when compared to previous models tested on a discourse coherence assessment dataset.

2 Related Work

2.1 EDUs and EDU Segmentation

EDUs represent the smallest assessable unit of a piece of text in this study. Slightly different from textual units like sentences, EDUs are discourse segments closely similar to constituents in a sentence syntax tree, shown as an example in Figure 1, which highlights by directional arrows the dependence of satellite EDUs on a nucleus EDU, and some of the connecting relations which they exhibit, such as an elaboration (elab) or attribution (attr) relation.

EDU segmentation involves extracting the start and end points of each EDU in the text. Initially treated as a syntactic parsing task due to the slight similarity of EDUs to clauses, neural approaches were later adopted utilizing a gold standard in discourse coherence datasets. Recent work such as that done by Lukasik et al. (2020) utilize encoder-decoder architectures to construe the problem as a segmentation-guessing task, which serves as a significant improvement in EDU segmentation from previous approaches such as those by Yu et al. (2019) and Lukasik et al. (2020).

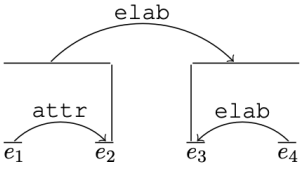
2.2 RST Parsing

Originally introduced by Mann and Thompson (1987), RST defines relations between two spans of text, namely a nucleus and a satellite. Each nucleus/satellite span is considered to consist of a single EDU. The idea behind this is that each body of text can be broken into such nucleus-satellite groupings (seen in Figure 2), with salient spans of text (nuclei) being independently interpretable, and linked to information only understandable with such a nucleus as pretext (satellites).

2.3 Discourse Coherence

Discourse coherence refers to the relationships between sentences that constitute everyday discourse or speech, and how intelligible they are when assessed as a whole. Discourse coherence maintains that real discourse is defined by coherence at both a local (paragraph) and global (structural arrangement) level. For example, there is generally more structure present in the layout of scientific paper when compared to impromptu speech in conversation, leading one to posit that the flow of ideas in the former may be understood more easily. Initially presented as a way of deconstructing and evaluating any text either written or transcribed, these studies require extensive linguistic knowledge and time-consuming analysis due to their highly qualitative nature. However, with the utilization of neural computation models, these formerly exhaustive processes of human evaluation are slowly becoming more easily accessible.

Local coherence is defined by the relationship between sentences in close proximity, the semantic similarities shared between them, as well as the salience of a discourse, or how they track the focus of discussion. These are highlighted as the systematic and topical ways in which clauses are related to each other at a local level. A way of measuring entity-based coherence, or how entities remain salient throughout discourse, was proposed by Grosz et al. (1995). This approach tracks which entities are forefront at different stages of a text by recording transitions between salient entities, firstly identifying their grammatical role in the text, shown in Table 3, before utilizing the entity grid model of coherence from Barzilay and Lapata (2008), seen in Figure 4, which shows early efforts of tracking the position and grammatical roles of



e_1: American Telephone & Telegraph Co. said it
e_2: will lay off 75 to 85 technicians here , effective Nov. 1.
e_3: The workers install , maintain and repair its private branch exchanges,
e_4: which are large intracompany telephone networks.

Figure 1: Example RST discourse tree, showing four EDUs, with nucleus/satellite relations indicated by directional arrows and labels.

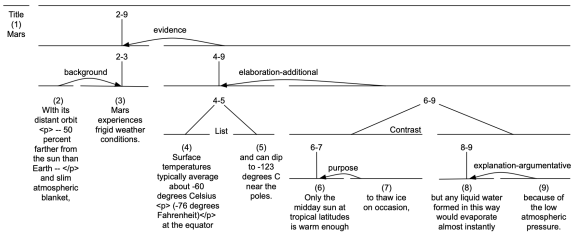


Figure 2: Example RST discourse tree, showing eight EDUs

1 [The Justice Department]$_s$ is conducting an [anti-trust trial]$_o$ against [Microsoft Corp.]$_x$ with [evidence]$_x$ that [the company]$_s$ is increasingly attempting to crush [competitors]$_o$.
2 [Microsoft]$_o$ is accused of trying to forcefully buy into [markets]$_x$ where [its own products]$_s$ are not competitive enough to unseat [established brands]$_o$.
3 [The case]$_s$ revolves around [evidence]$_o$ of [Microsoft]$_s$ aggressively pressuring [Netscape]$_o$ into merging [browser software]$_o$.
4 [Microsoft]$_s$ claims [its tactics]$_s$ are commonplace and good economically.
5 [The government]$_s$ may file [a civil suit]$_o$ ruling that [conspiracy]$_s$ to curb [competition]$_o$ through [collusion]$_x$ is [a violation of the Sherman Act]$_o$.
6 [Microsoft]$_s$ continues to show [increased earnings]$_o$ despite [the trial]$_x$.

Figure 3: Conversion of text to an entity grid representation, each cell indicates whether an entity is a subject (s), object (o), neither (x), or absent (-).

salient entities throughout a segment of text.

Global coherence involves the overall logical structure of a text, assessing how well it follows conventional discourse structures like scientific articles or stories. Studies on argument structure and scientific papers, such as those by Reed et al. (2008), Habernal and Gurevych (2016), and Memon et al. (2020) define argumentative relations and zoning to evaluate coherence. These studies provide foundational insights on the potential to identify topical and structural changes in textual discourse, but remain specialized studies in discourse-specific domains. Expanding the understanding of text structure for global coherence assessment is necessary for broader applicability.

2.4 Graph Neural Networks in NLP

We choose to employ a graph-based approach due to the highly-structural nature of assessing discourse coherence at a local and global level, discussed above in Section 2.3, and since our methods of graph construction (see Section 3.2) take both

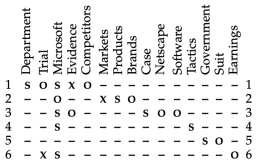


Figure 4: Discourse with entities marked and annotated with grammatical functions. (Barzilay and Lapata, 2008)

of these considerations into account.

Graph Neural Networks have gained popularity in NLP tasks due to their ability to model complex relationships between entities. Two prominent architectures are Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs), and they will be tested in this study.

2.4.1 Graph Convolutional Networks (GCNs)

GCNs, introduced by Kipf and Welling (2016), perform convolution operations on graph- structured data. They have been successfully applied to various NLP tasks, including text classification (Yao et al., 2018) and semantic role labeling (Marcheggiani and Titov, 2017). See Section 3.3.1 for a detailed explanation of the GCN implementation.

2.4.2 Graph Attention Networks (GATs)

GATs, proposed by Velickovic et al. (2017), introduce attention mechanisms to graph neural networks. This allows the model to assign different importance to different nodes in a node's neighborhood, potentially capturing more nuanced relationships in the data. See Section 3.3.2 for a detailed explanation of the GAT implementation.

3 Methodology

3.1 Datasets

The dataset used for this study is the Grammarly Corpus of Discourse Coherence (GCDC), with further information in Appendix Table A1, which includes texts from various sources such as Yahoo

forums, Hillary Clinton's emails, Enron emails, and Yelp reviews. Each text is a few paragraphs long and annotated with a coherence score ranging from 1 to 3, representing low to high levels of coherence. While the scoring system is not highly-nuanced, this dataset is particularly valuable because it provides a diverse range of discourse types, offering a robust basis for evaluating our models. We performed 10-fold cross-validation on each section of the dataset to ensure reliable and unbiased results.

3.2 Graph Data Construction

The data construction process involved several key steps to represent documents as graphs, in particular we use the subgraph and document-subgraph construction methodologies from Liu et al. (2023), however, in our approach, we construct the directed document graph and encode the information slightly differently, as explained in Figure 5 below.

- **Document Sentence Graph Representation**: Following Guinaudeau and Strube (2013), we represented documents as directed sentence graphs. Sentences were lemmatized, and cosine similarity scores of all noun pairs in each sentence were computed to form connections. For consistency, we used the same pre-trained GloVe embedding for comparing noun similarities. Sentences with a similarity score above a threshold were connected by directed edges, creating a graph representation of the document.

- **Feature Engineering for EDU Graph Representation**: Additional to sentence graphs, we used pretrained models for segmentation and parsing to create EDU graphs. Each text was segmented into EDUs using models from Lin et al. (2019), which typically results in shorter units than standard sentences. We then parsed these EDUs through a pretrained model for RST parsing (Lin et al., 2019). We avoid parsing any further since some non-coherent relations can be formed (an example is provided in Appendix Figure A1). As a result, quite a large number of EDU graphs are created, so we also create a separate dataset which creates links between nucleus-satellite heads based on the same similarity score mentioned above. We set the similarity threshold quite high (δ = 0.995) as to avoid over-connecting nucleus-satellite heads, and to retain the proper structural ordering of the text.

- **Subgraph Set Construction**: Each graph is represented as a subgraph set, which is a way to compare topological similarities between graphs (Shervashidze et al., 2009), and by extension a way to compare structural compositions of documents. We use Guinaudeau's (Guinaudeau and Strube, 2013) guidelines in defining a graph g is a subgraph of a graph G if the nodes in g can be mapped to the nodes in G and the connection relations within the two sets of nodes are the same. All subgraphs up to k-nodes are considered by enumerating all combinations of k-nodes and corresponding edges in G_i. As a result, all subgraphs with inter-sentence distances greater than some threshold w are filtered out since distant sentences are less likely to be related. We maintained a k-subgraph value of 4 and a maximum sentence distance of 8. As such, multiple subgraphs can have the same structure yet differ in node contents. The frequencies of all such isomorphic subgraphs are counted and used to represent a sentence graph as a k-node subgraph instead.

- **Doc-Subgraph Graph Construction**: A corpus-level undirected graph linking structurally similar documents via shared subgraphs was created. Edges in this graph indicate connections between subgraphs or between a document and a subgraph, weighted by subgraph frequency and inverse document frequency in the corpus.

3.3 Model Architectures

3.3.1 Graph Convolutional Network (GCN)

The baseline for comparison uses a GCN architecture based on Kipf and Welling (2016) to encode the doc-subgraph graph. GCNs perform operations on graph representations of data, learning node representations based on connectivity patterns and feature attributes. The convolution computation at each layer incorporates the adjacency matrix and degree matrix of the graph. Provided the graph input with $(N + M)$ nodes, Liu et al. (2023) define the convolution computation at the l^{th} layer as Equation 1:

$$H^{(l)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l-1)}\mathbf{W}^{l-1}) \quad (1)$$

Where $\tilde{A}$ is an adjacency matrix with self-connections created for each node, following Kipf

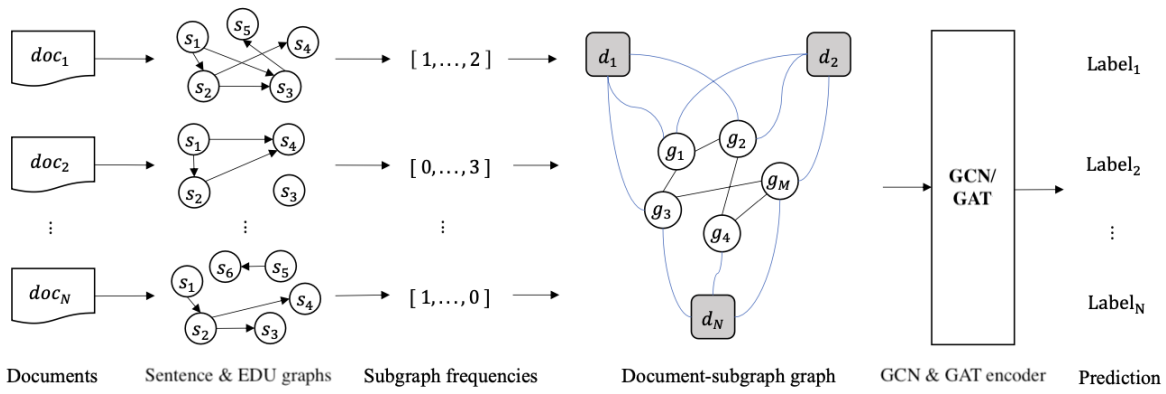


Figure 5: Overview of data processing method, with proposed changes made at document subgraph construction stage and encoder stage (Liu et al., 2023)

and Welling (2016), shown in Equation 2:

$$\tilde{A} = A + I_{N+M} \tag{2}$$

Where A represents that adjacency matrix and I_{N+M} an identity matrix. $\tilde{D}$ is the degree matrix and $\mathbf{W}^{(l-1)}$ is a layer-specific trainable weight matrix, with σ being a ReLU activation function.

The outputs are then fed into a softmax classifier which is expressed in Equation 3:

$$P = \text{softmax}(H^{(l)}) \tag{3}$$

The model is then trained by minimizing Cross-Entropy loss over document nodes, shown in Equation 4:

$$L_i = -\sum_{k=1}^{N}\sum_{j=1}^{C} Y_{i,j} \cdot \log(P_{i,j}) \tag{4}$$

Where N is the number of documents and C is the number of classes used in prediction.

3.3.2 Graph Attention Network (GAT) Implementation

We implemented a GAT architecture based on Velickovic et al. (2017), which incorporates attention mechanisms to learn node representations. GATs consider both graph structure and node feature attributes, allowing for more flexible parameterization. Our GAT model supports variable attention heads, layers, and other hyperparameters. In our implementation of the graph attention network, the attention mechanism is defined by Equation 5:

$$\alpha_{i,j} = \frac{\exp\left(\text{LeakyReLU}\left(\overrightarrow{a^T}[\mathbf{W}\overrightarrow{h_i}||\mathbf{W}\overrightarrow{h_j}]\right)\right)}{\sum_{k\in N_i}\exp\left(\text{LeakyReLU}\left(\overrightarrow{a^T}[\mathbf{W}\overrightarrow{h_i}||\mathbf{W}\overrightarrow{h_k}]\right)\right)} \tag{5}$$

Where $\cdot^T$ represents transposition and $||$ is a concatenation operation. When expanding to show the application of the LeakyReLU nonlinearity, we note that the negative input slope is provided by α, where smaller values will tend towards the standard ReLU function, whereas larger values will increase linearity for negative inputs.

Employing K multi-head attention results in the output feature representation for a multi-layer attention network calculated in Equation 6:

$$\overrightarrow{h_i'} = ||_{k=1}^{K}\sigma\left(\sum_{j\in N_i}\alpha_{ij}^k\mathbf{W}^k\overrightarrow{h_j}\right) \tag{6}$$

Where α_{ij}^k are normalised attention coefficients computed by the k^{th} attention mechanism and $\mathbf{W}^k$ is the corresponding weight matrix.

For the final prediction layer of the network, output features are represented by Equation 7:

$$\overrightarrow{h_i'} = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in N_i}\alpha_{ij}^k\mathbf{W}^k\overrightarrow{h_j}\right) \tag{7}$$

In which we average over the total number of attention heads K since concatenation is not feasible, and before any nonlinearity is applied.

Finally, we apply label smoothing and weighted cross entropy given by Equation 8 in order to aid model generalisation and reduce frequent overfitting found in early tests:

$$L_i = -\sum_{k=1}^{N}\sum_{j=1}^{C} w_j \cdot \Big((1-\varepsilon)\cdot y_{i,j}\log(p_{i,j}) + \frac{\varepsilon}{C}\Big) \tag{8}$$

Where the loss, L_i is minimised, w_j is the weight for the j^{th}class out of C classes and N doocuments, ϵ is a small positive value for label smoothing, $y_{i,j}$ is the true label for the j^{th}, i^{th} example in the smoothed class, and $p_{i,j}$ is the predicted probability for a given ${j_i}^{th}$ class of the k^{th} document, per standard Cross Entropy Loss calculation.

3.4 Optimization

We utilize the Optuna python library to automate and optimize the searching of the hyperparameter space. Due to computational constraints, we perform optimization on a single fold of each dataset for both GCN and GAT architectures. For the GAT hyperparameters, we search for the optimal combination of learning rate, hidden dimensions, dropout, weight decay, number of attention heads, and alpha. For GCN hyperparameters, we search for the best choice of hidden dimensions, graph convolutional layers, and learning rate. The optimal hyperparameters derived from the optimization search were applied to model training on the entire corpus.

3.5 Evaluation Metrics

Consistent with previous work, we use mean accuracy percentage as the main evaluation metric. We also consider F1 scores from each dataset to gain additional insights into model performance.

4 Results

4.1 Model Performances

Table 1 presents the average accuracies of the GCN and GAT architectures for each subgraph construction. As shown in Table 1, EDU preprocessing yielded higher accuracies for the GAT model across all datasets, with an average increase of 1.82 percentage points.

For the GCN architecture, the benefit of our methods on pure accuracy was less clear, per Table 1:

Our experiments revealed that the GCN architecture significantly outperformed the GAT model on average. Despite the potential for increased accuracy, the GCN model consistently outperformed the GAT model in our experiments. The highest-performing GAT trial achieved 60.15% accuracy on the Enron connected EDU dataset, which was still outperformed by a GCN architecture with fine-tuned hyperparameters.

Model	Subgraph	Average Acc
GCN	Sentences	61.23
	EDU	59.15
	Connected EDU	59.68
GAT	Sentences	52.87
	EDU	51.92
	Connected EDU	54.69

Table 1: GCN and GAT Subgraph Construction Comparison (Tuned Accuracies).

These results highlight the utility of our feature-extraction method using EDU segmentation and RST parsing, setting new performance benchmarks in discourse coherence modelling, while at the same time raising the important question of what sort of information contained in the corpus impacts the varying degrees of performance. In particular, what was it about the structure of the Enron corpus that elicited the most significant departure from previous benchmarks. This may be a question better answered either by analysis of more varied forms of discourse (mentioned in 5.1), or in being more selective with the length of the texts assessed in this investigation, such as using a sentence length filter condition like the one employed by Moon et al. (2019), especially considering that the global aspect of discourse coherence is very much a condition that takes into account information across the entire span of long-form discourse texts and documents rather than the shorter spans typical of the GCDC corpus.

Our runtime analysis revealed that the GCN architecture was significantly more efficient than the GAT architecture. GCN training averaged just below 1 second per epoch, while GAT training took between 1.5-1.9 seconds per epoch. This efficiency, combined with its strong performance, further justifies our recommendation of GCN as the more suitable architecture for this task.

4.2 Comparison with State-of-the-Art

Our method showed competitive performance across all GCDC datasets as seen in Table 2, where accuracy metrics of all previous approaches are shown, with current state of the art performances formatted in bold. Subscripts on some scores represent the value of 1 standard deviation.

Model	Yahoo	Clinton	Enron	Yelp	Average
(Li and Jurafsky, 2017)	53.50	61.00	54.40	49.10	54.50
(Lai and Tetreault, 2018)	54.90	60.20	53.20	54.40	55.70
(Mesgar and Strube, 2016)	47.30	57.70	50.60	54.60	52.55
(Mesgar and Strube, 2018)	$61.30_{0.84}$	$64.60_{0.89}$	$55.74_{0.90}$	$56.70_{0.78}$	59.59
(Moon et al., 2019)	$56.80_{0.95}$	$60.65_{0.76}$	$54.10_{0.89}$	$55.85_{0.85}$	56.85
(Jeon and Strube, 2020b)	$56.75_{0.83}$	$62.15_{0.88}$	$54.60_{0.97}$	$56.45_{0.97}$	57.49
(Jeon and Strube, 2020a)	57.30	61.70	54.50	56.90	57.60
(Liu et al., 2023)	$60.70_{1.03}$	$64.00_{1.36}$	$55.15_{1.14}$	$56.45_{0.94}$	59.10
(Liu et al., 2023)	$\mathbf{63.65}_{0.74}$	$\mathbf{66.20}_{0.81}$	$57.00_{0.81}$	$\mathbf{58.05}_{1.21}$	61.23
Our Method	$62.50_{1.25}$	$61.28_{1.68}$	$\mathbf{61.15}_{1.47}$	$56.53_{1.02}$	59.90

Table 2: Mean accuracy (std) results on GCDC.

Notably, we achieved state-of-the-art performance on the Enron dataset with 61.15% accuracy, outperforming the previous best of 57% (Liu et al., 2023).

4.3 F1 Score Analysis

Table 3 shows the F1-macro results for the dataset, comparing our method of EDU preprocessing - regardless of the level of subgraph connectivity - to the current state-of-the-art. As shown by the scores formatted in bold, our EDU preprocessing method consistently improved F1-macro results, establishing a new benchmark in the metric. However, these scores are still quite low and convey an issue in the evaluation of these datasets. The improvement in this metric yielded by our approach shows that deeper investigation is warranted to fully understand the degree to which graph constructions informatively reflect the content of the discourse they represent, and is necessary focus for future work.

This improvement in F1 scores is particularly important given the class imbalances in the GCDC dataset (examine Table 4 for the imbalance).

4.4 Error Analysis and Impact of EDU & RST Preprocessing

Our initial assumption was that a higher level of EDU subgraph connectivity and thus complexity of a text's subgraph representation would produce a direct benefit to how a document's inherent structure is encoded. Instead, we found that either construction method yielded an improvement in either F1-score or accuracy metrics. An example of the model and graph performances on both the Enron and Yelp datasets is shown in Tables 6 and 5, where the new benchmark values are formatted in bold.

Figures 6 and 7 show an analysis of confusion

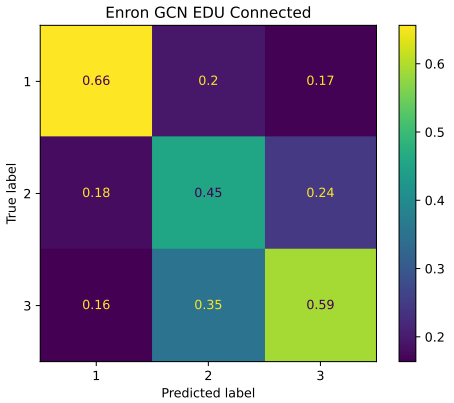


Figure 6: Enron GCN Connected EDU Confusion Matrices

matrices which revealed that across all datasets, the middle label (medium coherence level) was the most difficult to predict accurately, with a tendency to over-predict the high coherence label.

This suggests that while our representation doesn't yet comprehensively explain the graph structural representation of a text, our method of construction does elicit some important structural information from textual data. It also indicates that there may be an ideal degree of subgraph connectivity that can help the model better differentiate between coherence classes, which we consider grounds for future study.

4.5 Limitations of Baseline Models

We recognise that the pretrained models used for EDU segmentation and RST parsing from Lin et al. (2019), are comparable to state-of-the-art in the literature such as that by Lukasik et al. (2020) in their respective tasks, and still record competitive accuracies in their respective segmentation and parsing

Model	Yahoo	Clinton	Enron	Yelp	Average
Sentences	51.92	48.49	45.67	44.18	47.66
RST (Our Method)	**52.73**	**49.66**	**53.01**	**44.96**	**50.09**

Table 3: Mean F1 results on GCDC.

Dataset	Split	Label 1	Label 2	Label 3
Yahoo	Train	4560	1740	3700
	Test	820	410	770
Clinton	Train	2830	2060	5110
	Test	510	380	1110
Enron	Train	2990	1940	5070
	Test	620	500	880
Yelp	Train	2710	2180	5110
	Test	500	420	1080

Table 4: GCDC Dataset Label Counts.

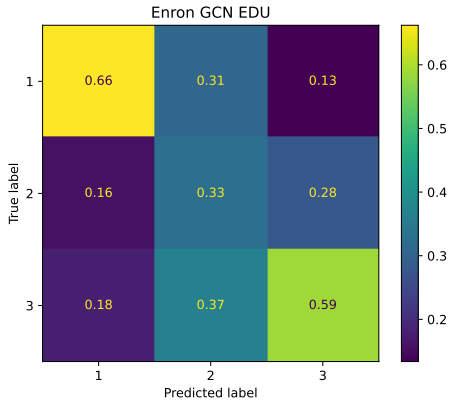


Figure 7: Enron GCN EDU Confusion Matrices

tasks, but show a lot of improvement to be made in those areas, meaning that it must not be overlooked that these accuracies can easily propagate and exaggerate any mistakes made in the data processing stages. In addition to this, the typical datasets of the RST Treebank and Penn Discourse Tree Bank used for training these tasks are quite dissimilar to the GCDC texts used. This leaves room for developing either shared datasets for the tasks or more rigorous pre-training of these models to suit the test data which could ultimately improve the fidelty of text subgraph representations.

4.6 Parameter Optimization Results

Parameter optimization results presented in Tables 5 and 6 show that the GCN model consistently outperformed the GAT model across various dataset constructions. This highlights the importance of careful hyperparameter tuning in achieving optimal model performance. We discovered there was great variation in the hyperparameters selected for the GAT model such as learning rate, attention heads and weight decay.

Model	**Untuned Acc**	**Tuned Acc**	**Highest F1**
GAT EDU	N/A	50.10	34.05
GAT Connected EDU	N/A	55.50	34.35
GAT Sentences	N/A	54.25	23.16
GCN EDU	59.40	58.93	46.86
GCN Connected EDU	59.33	61.28	**49.66**

Table 5: Clinton Optimization Results.

Model	**Untuned Acc**	**Tuned Acc**	**Highest F1**
GAT EDU	N/A	53.00	32.99
GAT Connected EDU	N/A	60.50	49.88
GAT Sentences	N/A	53.00	35.48
GCN EDU	58.92	60.13	51.28
GCN Connected EDU	59.60	**61.15**	**53.01**

Table 6: Enron Optimization Results.

Further, the variation seen in GAT hyperparameters was much greater than that of the GCN results, leading us to consider what the impact of a larger number of optimization tests would be adequate for this task, and thus highlight how considerations in identifying significant hyperparameters of the GAT architecture can reduce the search space and simplify its own optimization process.

5 Conclusion

This study has made several key contributions to the field of discourse coherence modelling:

1. We demonstrated that incorporating linguistic theory principles (EDU segmentation and RST parsing) has the potential to improve the performance of coherence modelling tasks, particularly in terms of F1 scores.

2. We established a new benchmark in accuracy on the Enron dataset of the GCDC corpus, and introduced a method of graph construction that improves F1-score across the entire dataset.

Our findings have several important implications:

1. The success of our EDU and RST-based feature extraction method validates the importance of incorporating linguistic theory into NLP models, and provides further direction for investigating how much information is properly conveyed in graph constructions using this method.

2. The superior performance and efficiency of GCN over GAT for this task suggests that simpler architectures may sometimes be more effective for certain NLP tasks.

3. The improvement in F1 scores across all datasets indicates that our method is particularly effective at handling imbalanced datasets, which is a common challenge in real-world NLP applications.

5.1 Future Work

While the GCDC dataset has typically been used as a benchmark dataset for evaluating discourse coherence, most samples are not truly long enough to emulate the length of what might be seen in free text generation. The TOEFL (Blanchard et al., 2013) dataset assesses coherence levels of much longer bodies of text than those of the GCDC dataset, and the findings from such a study would further aid in assessing the model's generalization to different types of text, since the TOEFL dataset contains 7 different prompts, meaning much more subject matter and thus textual content (semantic and structural) is included.

Additionally, a departure from typical accuracy metrics in a task with so few classes is warranted, and future work should aim to assess correlative performances against these classes instead.

Finally, while the use of LLMs was omitted in this study, it is recognized that useful insights may be gained in utilizing them for providing an additional point of comparison ranging from coherence score assessment to graph construction, and as such remains a focus for future studies.

6 Closing Remarks

By providing a more principled approach to representing text structure, we open new avenues for improving not only coherence modelling but potentially a wide range of NLP tasks that rely on understanding the structure and flow of text. As large language models continue to advance, the ability to evaluate and improve the coherence of generated text will become increasingly important. Our work provides a foundation for these future developments, bridging the gap between classical linguistic theory and cutting-edge machine learning techniques.

References

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv e-prints*, arXiv:1810.04805.

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

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.

Ivan Habernal and Iryna Gurevych. 2016. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43:125–179.

Yoichi Ishibashi, Danushka Bollegala, Katsuhito Sudoh, and Satoshi Nakamura. 2023. Evaluating the robustness of discrete prompts. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2373–2384, Dubrovnik, Croatia. Association for Computational Linguistics.

Sungho Jeon and Michael Strube. 2020a. 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.

Sungho Jeon and Michael Strube. 2020b. Incremental neural lexical coherence modeling. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6752–6758, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Lanlan Jiang, Shengjun Yuan, and Jun Li. 2021. A discourse coherence analysis method combining sentence embedding and dimension grid. *Complexity*, 2021:6654925.

Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR.

Thomas Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *ArXiv*, abs/1609.02907.

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.

Xiang Lin, Shafiq Joty, Prathyusha Jwalapuram, and M Saiful Bari. 2019. A unified linear-time framework for sentence-level discourse parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4200, Florence, Italy. Association for Computational Linguistics.

Wei Liu, Xiyan Fu, and Michael Strube. 2023. Modeling Structural Similarities between Documents for Coherence Assessment with Graph Convolutional Networks. *arXiv e-prints*, arXiv:2306.06472.

Michal Lukasik, Boris Dadachev, Gonçalo Simões, and Kishore Papineni. 2020. Text segmentation by cross segment attention. *ArXiv*, abs/2004.14535.

William C. Mann and Sandra A. Thompson. 1987. *Rhetorical Structure Theory: Description and Construction of Text Structures*, pages 85–95. Springer Netherlands, Dordrecht.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1506–1515, Copenhagen, Denmark. Association for Computational Linguistics.

Muhammad Qasim Memon, Yu Lu, Penghe Chen, Aasma Memon, Muhammad Salman Pathan, and Zulfiqar Ali Zardari. 2020. An ensemble clustering approach for topic discovery using implicit text segmentation. *Journal of Information Science*, 47:1–27.

Mohsen Mesgar and Michael Strube. 2016. Lexical coherence graph modeling using word embeddings. In *North American Chapter of the Association for Computational Linguistics*.

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.

Han Cheol Moon, Tasnim Mohiuddin, Shafiq Joty, and Xu Chi. 2019. A Unified Neural Coherence Model. *arXiv e-prints*, arXiv:1909.00349.

Chris Reed, Raquel Mochales Palau, Glenn Rowe, and Marie-Francine Moens. 2008. Language resources for studying argument. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Nino Shervashidze, SVN Vishwanathan, Tobias Petri, Kurt Mehlhorn, and Karsten Borgwardt. 2009. Efficient graphlet kernels for large graph comparison. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 488–495, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA. PMLR.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio', and Yoshua Bengio. 2017. Graph attention networks. *ArXiv*, abs/1710.10903.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv e-prints*, arXiv:2203.11171.

Yizhong Wang, Sujian Li, and Houfeng Wang. 2017. A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 184–188, Vancouver, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv e-prints*, arXiv:2201.11903.

Peng Xu, Hamidreza Saghir, Jin Sung Kang, Teng Long, Avishek Joey Bose, Yanshuai Cao, and Jackie Chi Kit Cheung. 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.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv e-prints*, arXiv:1906.08237.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2018. Graph convolutional networks for text classification. *ArXiv*, abs/1809.05679.

Yue Yu, Yilun Zhu, Yang Liu, Yan Liu, Siyao Peng, Mackenzie Gong, and Amir Zeldes. 2019. GumDrop at the DISRPT2019 shared task: A model stacking approach to discourse unit segmentation and connective detection. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 133–143, Minneapolis, MN. Association for Computational Linguistics.

A Appendix

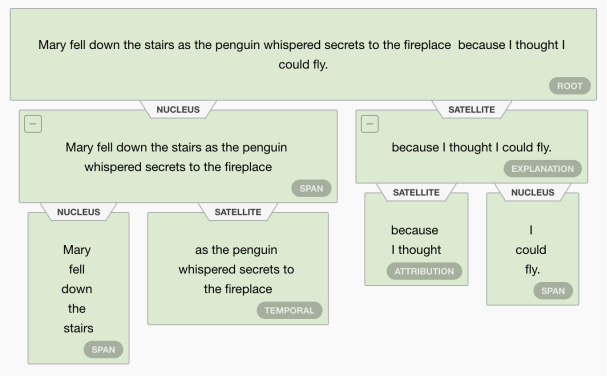


Figure A1: Example of how a completely nonsensical sentence will still be fully parsed, using model from Lin et al. (2019)

Dataset	Split	#Doc	Avg #W	Max #W	Avg #S
Yahoo	Train	1000	157.2	339	7.8
	Test	200	162.7	314	7.8
Clinton	Train	1000	182.9	346	8.9
	Test	200	186.0	352	8.8
Enron	Train	1000	185.1	353	9.2
	Test	200	179.1	340	10.1
Yelp	Train	1000	178.2	347	10.4
	Test	200	179.1	340	10.1

Table A1: GCDC Dataset Statistics. Doc, W, S refer to documents, words, sentences.