Towards Enabling Feedback on Rhetorical Structure with Neural Sequence Models

James Fiacco
Carnegie Mellon University
Pittsburgh, Pennsylvania
jfiacco@cs.cmu.edu

Elena Cotos
Iowa State University
Ames, Iowa
ecotos@iastate.edu

Carolyn Rosé
Carnegie Mellon University
Pittsburgh, Pennsylvania
cprose@cs.cmu.edu

ABSTRACT

Analysis of student writing, both for assessment and for enabling feedback have been of interest to the field of learning analytics. While much progress can be made through detection of local cues in writing, structured prediction approaches offer capabilities that are particularly well tailored to the needs of models aiming to offer substantive feedback on rhetorical structure. We thus cast the analysis of rhetorical structure in academic writing as a structured prediction task in which we employ models that leverage both local and global cues in writing. In particular, this paper presents a hierarchical neural architecture that performs this task. The evaluation demonstrates that the architecture achieves near-human performance while significantly surpassing state-of-the-art baselines. A multifaceted approach to model interpretation offers insights into the inner workings of the model.

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**; • **Applied computing** → *Computer-assisted instruction*; • **Human-centered computing** → Visualization techniques;

KEYWORDS

Rhetorical structure, neural sequence model, automatic essay evaluation, writing feedback, neural network interpretation, hierarchical, bidirectional LSTM, conditional random field

ACM Reference Format:
James Fiacco, Elena Cotos, and Carolyn Rosé. 2019. Towards Enabling Feedback on Rhetorical Structure with Neural Sequence Models. In *The 9th International Learning Analytics & Knowledge Conference (LAK19), March 4–8, 2019, Tempe, AZ, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3303772.3303808

1 INTRODUCTION

In the field of learning analytics, we desire to effectively and efficiently learn about the process of learning by observing learners, and language is one channel through which we can make those observations. Automated methods such as natural language processing and machine learning can be used as powerful tools to augment our ability to discern patterns in large quantities of data [30]. A full review of the wide range of language processing approaches that have been published in the field is beyond the scope of this conference submission, but have recently been reviewed in a learning analytics handbook [42]. The most closely related sub-area within learning analytics that our work relates to is automated essay scoring. Since the 60s, work in this area of research has turned from its initial focus on assigning a numerical score to an essay to providing actionable and interactive feedback [47]. Existing applications that provide feedback on writing are already in existence and offer feedback on elements of writing that are pervasive and can be detected by relatively local patterns [31, 53]. Building on and extending this past work, we propose a method that will allow for new, previously unavailable automated feedback that can be triggered in response to the rhetorical structure of a document, as operationalized as its rhetorical moves and steps [4].

With the tremendous amount of recent attention placed on neural architectures in virtually every area of language technologies and natural language processing, highly sophisticated structured prediction models have been published in areas such as parsing [13, 27], named entity recognition [24], and part-of-speech tagging [16]. It is therefore attractive to approach the detection of the rhetorical structure of writing as a structured prediction task as well, due to both the fine-grained nature and context sensitivity of the distinctions it involves. Neural models are non-linear (i.e., multilayer) weight-based machine learning models trained through an algorithm known as backpropagation, and though there has been a resurgence of interest in this approach recently, the methods have been in use for decades [43].

Some popular theories underlying analysis of the rhetorical structure of writing, such as Rhetorical Structure Theory [52], offer complex and nuanced representations at the discourse level. Nevertheless, recent tasks for rhetorical structure analysis in the language technologies domain have simplified this problem by focusing on very local judgments, and high performance has been mainly achieved on coarse-grained distinctions [18, 26, 54]. Discourse-level processing has targeted problems such as shallow discourse parsing and sentiment analysis. In this paper, we present a hierarchical neural architecture capable of achieving near-human performance for annotator agreement on a corpus annotated with fine-grained rhetorical relations, which are textual relations that are based in genre analysis. These structures require sensitivity to context for accurate detection. The evaluation demonstrates that this approach substantially beats baseline approaches on this task.

Neural networks, and more specifically recurrent neural networks (RNNs), have repeatedly been shown to have the ability to solve complex sequence modeling tasks at both the sentence level

LAK19, March 4–8, 2019, Tempe, AZ, USA

ACM ISBN 978-1-4503-6256-6/19/03…$15.00
https://doi.org/10.1145/3303772.3303808

Moves	Number of Words	Number of Units	Percent
Establishing the territory (I-M1)	446140	18284	7.22
Identifying a niche (I-M2)	117934	4844	1.91
Addressing the niche (I-M3)	139886	5511	2.18
Contextualizing the study methods (M-M1)	284527	20279	8.00
Describing the study (M-M2)	1130217	69997	27.63
Establishing credibility (M-M3)	139857	6201	2.45
Approaching the niche (R-M1)	386083	17715	6.99
Occupying the niche (R-M2)	916557	45142	17.82
Construing the niche (R-M3)	245141	10610	4.19
Expanding the niche (R-M4)	19962	810	0.32
Re-establishing the territory (D/C-M1)	689095	29515	11.96
Framing the new knowledge (D/C-M2)	395000	15248	6.02
Reshaping the territory (D/C-M3)	70689	2533	1.00
Establishing additional territory (D/C-M4)	173992	6617	2.61
Total	5155080	253306	100.0

Table 1: Corpus statistics.

[6] and the document level [49, 55]. While these results have practical value, the scientific value of the work would be enhanced if the resulting models were more interpretable. Though advances in interpretability of network architectures in areas like computer vision have inspired similar efforts in the language technologies field [25], progress has been slow, especially in regard to interpreting recurrent neural models [2, 21, 23, 28, 39, 40, 50]. Our intuition is that, due to the highly dynamic and complex nature of information processing within neural architectures, each single technique for interpretation offers only partial insights. In this paper, therefore, we take a multifaceted, integrative approach involving multiple interpretation methods working together to provide a detailed analysis. The insights gained through such an analysis can allow for previously unavailable feedback to writers. Thus, our contribution of this paper above the specific results in rhetorical analysis of academic writing is an illustration of an integrative model interpretation approach that offers visibility into the context sensitive processing and that allows for automated writing feedback that in future work can be applied to interactive educational systems.

2 RHETORICAL ANALYSIS OF STUDENT WRITING

In order to test our modeling and interpretation approach, we use two datasets, the Research Writing Tutor (RWT) corpus and the Intelligent Academic Discourse Evaluator (IADE) corpus. Both of these dataset are well suited for our goals as each are annotated at a fairly granular level with rhetorical moves and steps as described by Swales [51].

2.1 Research Writing Tutor (RWT) Corpus

We used the RWT corpus which comprises 900 research articles published in high-impact peer-reviewed journals in 30 academic disciplines [9]. Each discipline is represented by 30 articles that report empirical studies and were authored by different researchers. The disciplinary sub-corpora were compiled following the criteria outlined by Sinclair [48]. The journals and articles were selected in collaboration with faculty in respective fields. The size of the corpus is 5,155,080 words (Introduction 703,960 words; Methods 1,554,601 words; Results 1,567,743 words; Discussion/Conclusion 1,328,776 words).

Rhetorical Structures: The analysis of the RWT corpus centers on rhetorical constituents defined in the tradition of genre theory [51]. Swalesian genre theory is grounded in English for Specific Purposes (ESP), a field within Applied Linguistics. From this theoretical perspective, genres are text types that enact particular types of social practices of particular discourse communities. As text types, they are viewed as dynamic rhetorical structures that strategically package content in ways that reflect the norms and expectations of discourse communities [4].

These structures represent the rhetorical conventions of genres, which in ESP are conceived of as communicative goals, termed *moves*, and various functional strategies, termed *steps*, which writers invoke in their writing in order to accomplish the social and rhetorical goals of the moves. In earlier work, the RWT corpus was used to devise a cross-disciplinary move/step framework for the Introduction, Methods, Results, and Discussion/Conclusion (IMRD/C) sections found in academic writing [9–11]. This framework was used to manually annotate the entire corpus, which we use for our evaluation.

Data Annotation: As described in Cotos et al. [9, 11], a coding protocol was developed prior to corpus annotation, which contained guidelines and examples for annotation training. According to the protocol, the unit of annotation was defined as a functional segment of text. Because rhetorical functions are often intertwined, the unit of annotation was first considered at the sentence level. A sentence that had one communicative goal realized with one functional strategy was assigned one move and one step. A sentence that had one communicative goal realized with more than one functional strategy was assigned one move and several respective steps of that move.

Some sentences contained elements that had more than one communicative goal; in such cases, the full sentence was tagged with a primary move and step, and secondary move/step tags were assigned to the multi-functional segments of that sentence. This annotation approach thus captured nested rhetorical functions, rendering a more complex dataset compared to previous works where corpora were coded only for salient functions [20].

Overall, the annotated data amounted to 253,306 units, with individual sections consisting of: Introduction 28,639 units; Methods 96,477 units; Results 74,277 units; Discussion/Conclusion 53,913 units. Table 1 summarizes the distribution of moves for each section. These unit numbers include both primary and secondary functions. Two measures of reliability indicate satisfactory agreement rates for this challenging multi-functional annotation task. For moves, Cohen's Kappa (κ) for pairs of annotators ranged between 0.60 and 0.99, and Intraclass Correlation Coefficient (ICC) estimates of reliability among the three coders was 0.86. For steps, κ ranged between 0.59 and 0.81, and ICC = 0.80.

2.2 Intelligent Academic Discourse Evaluator (IADE) Corpus

In order to connect our work more strongly to previous work, we also evaluate our model with the Iowa IADE corpus, on which Cotos and Pendar [12, 35] have the state- of-the art performance record. This corpus is much like the RWT corpus only simpler. In particular, it only includes 3 moves, 17 steps, and only contains introduction sections. Because sections other than introductions are not included, this dataset is substantially smaller, so in our evaluation we primarily focus on the RWT corpus except to compare to previous work.

3 RELATED TECHNICAL WORK

This paper presents an applied, technical approach in the hybrid structural prediction and neural sequence modeling paradigm, which has experienced a recent resurgence of interest. Within that sphere, this work additionally draws on the neural network interpretation and discourse analysis fields. We therefore begin with a broad discussion of how our work positions itself with respect to discourse and rhetorical analysis, specifically relating to current standard datasets and tasks. Next, we describe prior work in neural sequence modeling with structured output spaces. We will further explain how a structured approach allows us to examine the learning of the algorithm. Lastly, we transition into a discussion of neural network interpretation, as we approach our method with a priority of at least partially explaining the complex cfunction computed by the neural model in order to perform this task.

3.1 Discourse and Rhetorical Analysis

Work in the field of Language Technologies related to analysis of rhetorical structure has largely focused on two main corpora. First is the Penn Discourse Treebank (PDTB) [32], which was a SIGNLL Conference on Computational Natural Language Learning (CoNLL) shared task in 2015 [54]. The second is the Rhetorical Structure Theory (RST) Discourse Treebank [5], which is smaller and has been the focus of less computational work. Much work on the former dataset has focused on a coarse-grained prediction of implicit discourse relations and shallow discourse parsing [17, 18, 45]. The PDTB was created as a low-level discourse modeling corpus, generally focusing on local relations within a sentence or in adjacent sentences.

In contrast, RST and its related corpus are designed as a hierarchical structure from Elementary Discourse Units to full documents, modeling complex dependencies therein [52]. The relatively small size of this corpus has discouraged work exploring the capabilities of neural architectures using it as a data source. Because of that, this paper leverages a corpus produced within the field of Corpus Linguistics with the purpose of providing a basis for offering automated feedback on student writing, namely the Research Writing Tutor (RWT) Corpus [9].

3.2 Neural Sequence Modeling and Structured Prediction

State-of-the-art results in sequence modeling often use the recurrent neural network (RNN) conditional random field (CRF) as described in Huang et al. [16], a method of layering a CRF over a recurrent network in order to model tag transitions over a whole sequence [19, 29, 33]. Related to our work is a method for a two tiered recurrent model to perform discourse mode detection for the purpose of essay grading [49], which we extend by including the CRF layer from Huang et al. [16] to stabilize training.

Further improvements to sequence prediction models can be achieved using a method called dual decomposition [44]. This is a technique for jointly finding the best scoring sequence of predictions of two related tasks. It leverages the relatedness of each task to enforce a constraint that both sequence predictors must agree on compatible sequences. This technique was used to enforce agreement between our rhetorical move predictor and our rhetorical step predictor.

3.3 Interpreting Neural Networks

The topic of interpreting neural networks has recently come to the forefront of the field as neural models continue to demonstrate their predictive power by advancing nearly every area of Language Technologies [3, 38, 46]. In our view there are three motivating questions that are central to the successful interpretation of a neural model:

- Which upstream inputs influence a neuron's activation level, and conversely, how do its activation levels affect downstream task performance?
- To what degree is activation related to input at a single time step as opposed to being context-sensitive (i.e., demonstrating influence from multiple time steps)? This is challenging to determine within recurrent neural models as there is a large amount of interdependence between neurons both within and across time steps.
- Which neurons carry the most statistical influence over the final classification? A principal challenge of neural network interpretation is that there are too many neurons to attempt understanding what each contributes to the big picture. Narrowing to those that are most important would make interpretation more tractable.

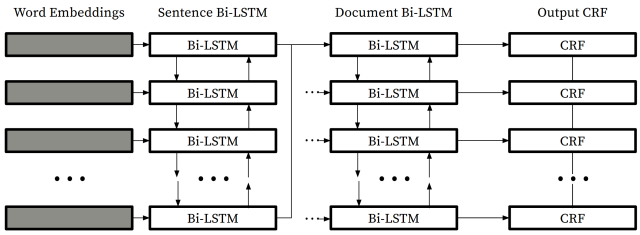

Figure 1: Tiered Bidirectional LSTM with CRF.

Out of the sample of work that has influenced our qualitative methods, we found that previous work has primarily focused on the first of these questions. Such work follows: Karpathy et al. [21] proposed a way of probing into recurrent models by tracking activations across inputs and used human intuition to interpret the patterns observed, which is expanded upon by Strobelt et al. [50]. Furthermore, work has been done at interpreting neural models through use of simpler classifiers to explain mistakes made by the more complex model [23, 40]. Linear classifier probes have also been introduced by Alain and Bengio [2] to co-train simple linear models to illustrate functions that particular layers are performing in arbitrarily deep models. The predictive capacity of individual neurons has also been explored in Radford et al. [39] and Lin and Wu [28] to show how a sampling of neurons influence the output of the network.

Our work on interpretation expands on this prior work while simplifying some procedures and introducing some of our own. Our goal is to address all three questions with an integrative approach, making use of multiple lenses and then integrating the disparate pictures into a unified vision of network function.

4 CONTEXT-AWARE RHETORICAL STRUCTURE LABELING

4.1 Model

Following previous work in discourse sequence labeling [49], we use a hierarchical neural architecture consisting of an embedding layer, a sentence-level recurrent layer, and a document-level recurrent layer. Given that rhetorical structures have strong correlations with global contextual structures, we also include a CRF layer [16] as the output layer of the document level recurrent network. This allows the model to leverage the powerful recurrent layers to learn the more subtle differences between the types of rhetorical structures, rather than simply learning the transitions between different tags, which could as easily be handled by the CRF layer alone. Altogether, this approach gives us the four-layered structure illustrated in Figure 1: a word embedding layer, a sentence level recurrent layer, a document level recurrent layer, and an output CRF layer. For a technical description of the mechanics of the model, see Huang et al. [16].

Word Embeddings: At the base of the model, we have a word embedding layer that is initialized with a pretrained distributed representation of the text referred to as GloVe embeddings [36]. This representation can be thought of as analogous to what might be obtained using Latent Semantic Analysis. Over the course of training, the embeddings are fine-tuned to the dataset through the neural training algorithm, namely backpropagation of errors.

Sentence Bi-LSTM: Each sentence was encoded to a single vector with a neural architecture referred to as Bidirectional LSTM [14] by feeding each word embedding sequentially into two separate LSTMs, one forwards and one backwards. The final state of each LSTM is then concatenated together to produce the full sentence embedding. This layer models the sequential information within the sentence with the resulting vector representing the most salient features detected within the sentence. The same sentence level Bi-LSTM is applied to each sentence in the document separately and the vectors are aggregated into a sequence of sentence vectors for the document level Bi-LSTM.

Document Bi-LSTM: This layer models the sequential information between sentences within the document. In this layer, sentence embeddings are sequentially fed into both the forward and backward LSTMs, and at each time step, the network state is outputted. The output of this layer is the concatenation of the forward and backward states at the respective time step in each network for each sentence. In this way, we obtain new sentence embeddings that not only encode the most important sentence level features, but also inter-sentence information.

Output CRF: The CRF layer models the transitions between tags in the resulting sequence prediction. This allows us to use the predictions for past and future tags to influence the prediction of the current tag. This also offloads the task of modeling sentence tag ordering from the document level Bi-LSTM, allowing it to focus on the linguistic information contained in the input. The CRF layer takes in the sentence embeddings at each time step in the network and predicts a probability for each tag for that time step and the transition probabilities for each tag to another. These probabilities can then used to compute the most probable sequence of tags.

5 INTERPRETATION METHODS FOR MODEL

To answer our motivating questions from the Interpreting Neural Networks subsection and to build a understanding of our model, we use six easy qualitative analysis techniques, implemented in the open source machine learning package called Scikit-learn [34].

5.1 Qualitative Methods

Finding the Most Salient Neurons: We are most interested in how the structure of learned activations of the model reflects insights about the data that pass through it. For each tag-neuron pair, we compute a χ^2 test for independence. We then rank the neurons for a given tag on their test statistic and take the most connected set of neurons as our candidate set for further evaluation in terms of a given tag.

Approximation of Activations and **Context Contribution:** To determine the approximate contribution of content versus context to a neuron's activation, we make linear approximations of each of those influences on the neuron. For content, we approximate each neuron's activation using what is known as a Ridge regression [41] on a simple representation of the text referred to as a bag-of-words feature representation, we get a window of insight into how intra-sentence content words contribute to neuron activation.

Concretely, to determine the degree to which context influences the activation, we fit a Ridge regression model to predict an activation state based on the other, same-numbered neuron activations before or after it (depending on whether it is a forward or backward neuron) within the document. We can investigate the of shape of contextual contributions over time by plotting the regression weights of sentences in a relative position to the neuron of interest, normalized by the frequency of document lengths. For both regressions, the strength of the correlation indicates relative influence of either the content or context over a given neuron.

Activation Heatmaps: The most familiar visualization of neural networks must be the activation heatmap. We use it to gain insight into how neurons are influencing the the network output. We break up our activation heatmap to reflect the average activations for each correct rhetorical step tag. These averages are colored to represent a negative or positive valued average activation. Using knowledge of rhetorical structures, we can order the classes so that they are grouped by rhetorical moves and document section which allows patterns to be more readily seen.

Document Colorization: To verify our interpretations from previous methods, we turn to the data, augmented with another familiar visualization technique explored in Karpathy et al. [21] that colors the items in the sequence based on the values of the activated neuron. We visualize this in two ways: the first is coloring each sentence of the actual text of the document based on the correct tags and the second is coloring just the sequence of correct tags. We found that both of these techniques were important and complementary as the first technique allows the researcher to examine the neurons in context of the input space while the latter allows the researcher to quickly see any neuronal structure in the output space.

Single Activation Linear Classification: The last analysis technique is also one that has been used to great effect in prior work [39] and can be used to further evaluate how much influence an individual neuron has on the final classification. We thus make binary classifications for each class given a single neuron's activation as input. We can then compare this score with that of the full model.

5.2 Implementation Details

We implemented the multi-tiered Bi-LSTM model using the Keras deep learning library [7] with the TensorFlow tensor library [1] as a backend. The sentence level Bi-LSTM had 512 hidden units (256 forward, 256 backward) and a 0.5 dropout rate. The document level Bi-LSTM had 128 total hidden units (64 forward, 64 backward) with a 0.5 dropout. For the CRF layer, we used Viterbi decoding [16]. Our learning rate was 0.001 for the Adam optimizer [22]. For efficient mini-batching, we padded sentences to a max sentence length of 75 words and we padded the document to a max length of 200 sentences. For padded sentences, we introduced a null class prediction and discarded the predictions from padding sentences for final output, truncating the sequence of outputs at the number of sentences in the actual document.

Our word embeddings were fine-tuned 300 dimensional pretrained Global Vectors for Word Representation (GloVe) embeddings [36]. These vectors were pretrained over the Common Crawl[1]. Unknown words were assigned one of 100 random vectors.

Choices for hyperparameters and the numbers of neurons in the constructed models were the default values and a standard number for sentence representations respectively. The priority was to test the modelling assumptions of the model architectures rather than fine tuning the models for optimum performance. It is possible that performance could be increased marginally by using model tuning techniques or ensembling [37].

6 EVALUATION

6.1 Training Tasks

The RWT Corpus provides two natural prediction tasks with which we can use to evaluate the performance of the neural sequence model and then apply our model interpretation approach to the resulting model. We have the fine grained rhetorical step prediction task and we have the more coarse grained move prediction task. Furthermore, we can make these two predictions jointly, thus leveraging the relatedness of steps and moves to improve performance of the resulting joint model on both. The dataset was divided by randomly holding out 10% of the documents for the test and using the rest for train.

Step Prediction: For our initial set of experiments on this dataset, we chose to address the fine-grained task of rhetorical step prediction. All models treat the task as a multi-class prediction problem. We include the results from the previous state-of-the-art classifier from Cotos et al. [8] to demonstrate the difficulty of this task. We evaluate using the standard precision, recall, and f-measure as well as Cohen's kappa, which accounts for the unbalanced class sizes.

Move Prediction: We also trained our model to predict the rhetorical moves in the dataset. This is a more coarse grained prediction task as compared to the step prediction task described above. We were primarily interested in this task for the potential to improve the performance of the more fine grained task, by making predictions at the coarse level.

Joint Move and Step Prediction: Because of the related nature of moves and steps in the rhetorical analysis paradigm that we are using in this work, we also experimented with dual decomposition [44] as a method of integrating our move prediction model and our step prediction model. This method forces each model to produce a move or step that is consistent with the prediction of the other model. For example, many steps only occur as a part of a single type of move. If our move predictor guessed a move and our step predictor guessed a step that only occurs under a different move, we know that at least one of them is wrong so we penalize their choices until they agree. This penalty step is only performed at decoding time in the CRF layer, that is it is not a fixed set of parameters learned along with the model, rather it is a sub-optimization problem solved independently for each data instance.

[1] http://commoncrawl.org/

Model	Cohen's Kappa	Precision	Recall	F1
SVM [12]	0.441	0.46	0.48	0.46
MaxEnt [8]	0.452	0.47	0.49	0.46
Sentence Bi-LSTM Only	0.531	0.57	0.63	0.58
Two-tiered Bi-LSTM Only	0.110	0.13	0.24	0.13
Word Embeddings w/ CRF*	0.723	0.75	0.74	0.74
Sentence Bi-LSTM CRF*	0.709	0.74	0.73	0.73
Two-tiered Bi-LSTM CRF[2]	**0.751**	**0.77**	**0.77**	**0.76**

Table 2: Step classification results on RWT Corpus.

6.2 Baselines

We want to evaluate two conditions: the effect of adding a document level CRF to aid in classification, and the effect of the individual model architecture. We use baselines that evaluate these conditions and connect our evaluations with previous work. The models that we evaluated follow:

Support Vector Machine (SVM): This is the current state-of-the-art classifier on the IADE dataset [12]. It consists of an SVM [15] that predicts the rhetorical move based on primarily N-Gram features then, with that prediction and the same set of features, predicts the rhetorical step.

MaxEnt Ensemble: This baseline [8] is composed of an ensemble of six maximum entropy classifiers. Each classifier has features of either unigrams, bigrams, or trigrams. For each n-gram size, there is a classifier that uses n-grams of stemmed text and one that uses n-grams of part-of-speech tags. In our implementation of these baselines, this model performed better than the SVM baseline above on the larger RWT dataset. For this reason we primarily compare our models with this one.

Sentence Bi-LSTM: This baseline contains just the sentence level Bi-LSTM as described above with a softmax output layer to classify the segment.

Two-tiered Bi-LSTM: This baseline is comprised of the two-tiered Bi-LSTM structure described above with a softmax output layer for each sentence rather than the CRF.

Word Embeddings with CRF: This baseline flattens all the word embeddings for the sentences in the document and then passes them through the CRF layer to predict classes.

Sentence Bi-LSTM CRF: This baseline has the CRF layer directly above the sentence level Bi-LSTM, without the document level Bi-LSTM.

7 RESULTS AND ANALYSIS

7.1 Quantitative Performance

As can be seen in Table 2, the two-tiered Bi-LSTM model with the CRF performed significantly better across all metrics than the next best model. It is interesting to note, however, that the two-tiered Bi-LSTM without the CRF performed far-and-away the worst. A qualitative analysis of the output of that model showed that it had a tendency to latch onto common classes and would not predict the less common classes in a way that the models with the CRF did, a behavior that was present to a lesser extent in the sentence level Bi-LSTM without the CRF. In all cases, adding the CRF to the output layer of the network increased performance dramatically. This can likely be attributed to its role in guiding the model to only output likely sequences of tags. It reinforces that these rhetorical structures are highly dependent on one another and, even though certain classes can be highly uncommon, they have a specific place in the document that can be inferred by the CRF layer through features extracted by the Bi-LSTMs. Furthermore, given that the two-tiered Bi-LSTM CRF performed better than the strictly sentence level Bi-LSTM CRF, we can also see that contextual information from other sentences in the document can be utilized to increase classification performance.

	Cohen's Kappa	F1
Step Prediction Only	**0.751**	**0.76**
Step with Dual Decomposition	0.722	0.74
Move Prediction Only	0.778	0.80
Move with Dual Decomposition	**0.785**	**0.81**

Table 3: Performance of Move, Step, and Joint sequence predictions of Two-tiered Bi-LSTM with CRF.

From our experiment with joint training we noticed an unexpected behavior. Namely, the performance of our step predictions was degraded by about 2% while our move prediction performance was improved by about 1%. This may have occurred because the dual decomposition algorithm can potentially favor the more confident model in the process to agree on most likely sequences. Because there are fewer classes for the move predictor it may have been generally more confident in its predictions (whether correct or not). This would lead the resulting agreed upon jointly predicted sequence to favor the predictions made by the move model rather than the step model. Thus the only time when the move model would predict a different sequence via dual decomposition would be when the step model was very confident that it was correct and had chosen a step that did not fall under the predicted move. This would reasonably result in the move prediction performance to increase. Conversely, the step model would be forced to change its prediction whenever it disagreed with the move model and was not especially confident leading it to share more of the mistakes of the other model without benefiting from it as much. It may be possible to control for this imbalance in future work.

From Table 4 it appears that the neural model performs worse on the task than the simpler models. This is likely because the IADE

[2]Significant improvement over baselines, $p < 0.001$

	Precision	Recall	F1
SVM (Cotos et al. [8])	0.69	0.59	0.61
MaxEnt	0.64	0.64	0.61
Two-tiered Bi-LSTM CRF	0.57	0.63	0.58

Table 4: Step prediction on IADE Corpus.

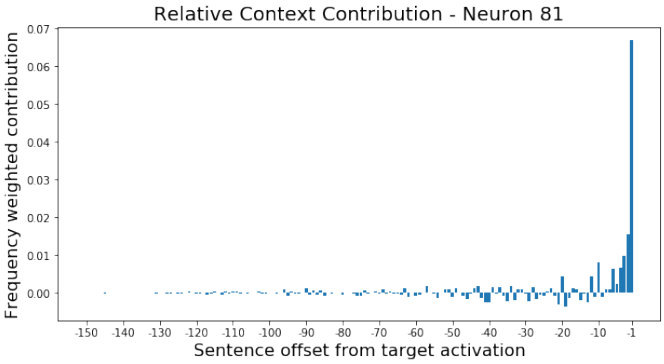

Figure 2: Neuron 81 normalized context approximation weight indexed by offset to predicted activation. This is a forward LSTM neuron so offsets correspond to relative index of context word.

corpus is substantially smaller than the RWT corpus as it only contains the introductions to around 600 documents (compared to the approximately 1000 documents with introductions, methods, results, and discussion in the RWT corpus). This does highlight the need for sufficient data for this approach. However, given the performance gains when scaled up to the larger corpus, it is a reasonable trade-off.

7.2 Qualitative Analysis

We now demonstrate the value of our multi-lens interpretation approach by comparing two steps via our learned model activations and evaluate it qualitatively against the descriptions in the annotation guide for the data set.

Finding Neurons to Analyze: We focused our neural structure analysis on the output of the document level Bi-LSTM before the CRF layer. This was chosen as it should contain the richest data, including both sentence and document level features pertinent to our classification problem. In principle, these techniques could be applied to other layers as well.

The steps that we compare in this paper were selected by computing the set of ten most salient neurons for each step via the χ^2 analysis described in Qualitative Methods section. We chose the pair of steps that had the smallest Hamming distance between the sets of salient neurons. This enables us to focus on a pair of steps that use similar neurons, allowing us a principled way to analyze a smaller set of differentiating neurons for the paper's illustrative purposes.

Through this method, we select the steps *Justifying Specifics* and *Accounting for Results*. Out of the five overlapping highly salient

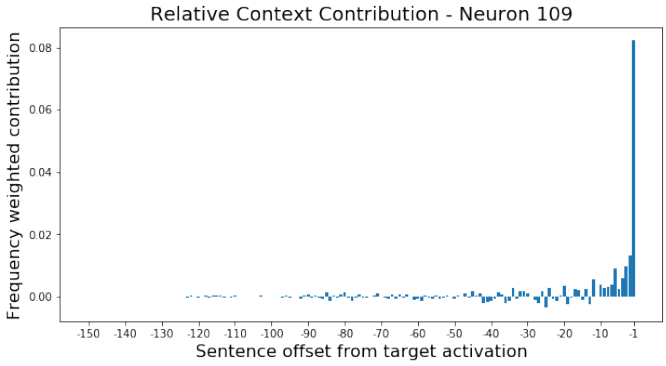

Figure 3: Neuron 109 normalized context approximation weight indexed by offset to predicted activation. This is a forward LSTM neuron so offsets correspond to relative index of context word.

neurons, we found that neurons 81 and 109 were the most differentiating based on the single neuron classification test from the Qualitative Methods section. From Cotos et al. [9], *Justifying Specifics* describes providing reasoning and importance of choices that may influence the results of an experiment with the intent to provide a scaffold for the presentation of the results. *Accounting for Results* occurs with the intent to establish the meaning of observed results by underscoring their soundness and anticipating questions from a reader. We would therefore expect to see differentiating characteristics of the first step be language establishing an experiment that will be described, while the other should be distinguished by language that relates to the results of a given experiment.

Characterizing Neuron Specifics: Using the heatmap in Figure 4, we examine the commonalities between the neurons with other steps, not just the two we are interested in. Neuron 81 is commonly higher than average in steps regarding gaps or problems in existing work, and proposing ways to fix them. The set of unigrams that best predict the level of this neuron's activation, Table 6, have highest weight around words that appear in section headings and lowest weights on specific numbers and numerals indicating a section heading. This may indicate that this neuron is capturing information regarding the structure of the document as referenced in non-heading sentences; this is consistent with the function of *Justifying Specifics* as a step-to-scaffold presentation. This can further be seen in the colored documents where the neuron tends to be high between steps similar to *Stating Specific Results*. It is evident that contextual values of this neuron explain some of the variance in its activation from Table 5 and Figure 2.

Similarly, we analyze Neuron 109 to see that positive activations co-occur in steps talking about present research and steps talking about the research that was done in the paper, while notably not activating on steps related to study specifics. Given the definition of our steps of interest, this seems unexpected. However, the unigram regression model in Table 6 shows that the highest weighted words are abstract words such as "summarizing" and "substantiate", which are consistent with *Accounting for Results*, while specific nouns that likely appeared in the study's specifics are down-weighted.

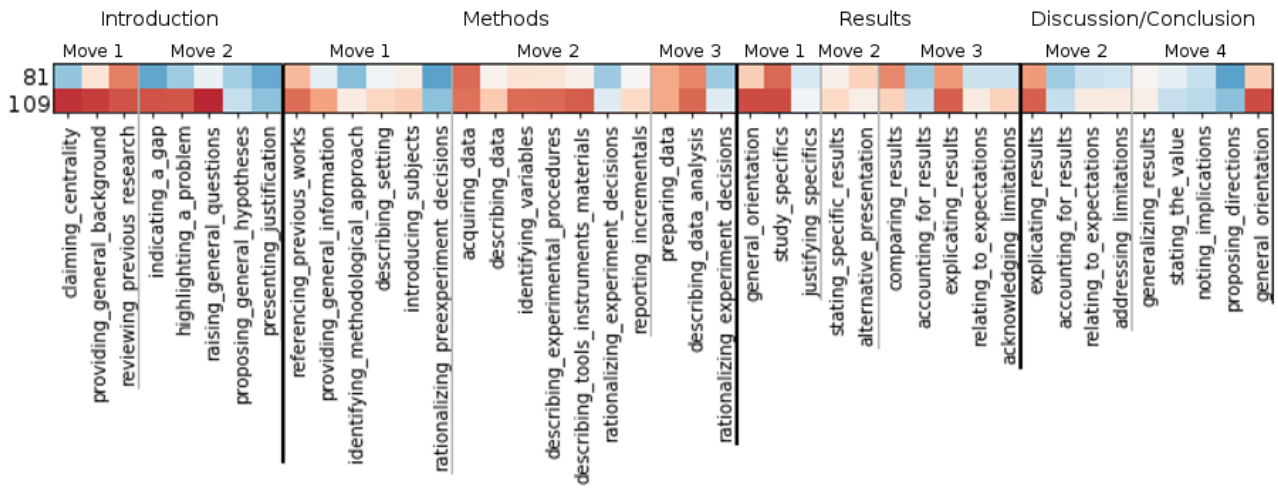

Figure 4: Document level Bi-LSTM neuron activation heatmap. Blue indicates above average positive activation for a given tag over the data, red indicates below average negative.

Coloring the data as in Karpathy et al. [21] does not reveal notable structure on its own, though with the other methods, the pattern that the unigram model demonstrates appears to hold.

While each of these specific neurons seem to carry small amounts of information from contextual sources in the document level Bi-LSTM, we can also see that distinguishing between the tags has strong document level contextual dependencies via the CRF layer; see Table 2. In the CRF transition weights, we can see this structure where *Accounting for Results* has strong weight near steps *Announcing Principal Outcomes* and *Reporting Incrementals* while *Justifying Specifics* has strong weight near *Relating to Expectation*, *Generalizing Results*, and *Comparing Results*.

	Unigram R^2	Context R^2
Mean	0.893	0.355
Stddev	0.018	0.273
Max	0.925	0.968
Min	0.849	0.012
Median	0.896	0.280
Neuron 81	0.855	0.109
Neuron 109	0.874	0.159

Table 5: Residual sum of squares statistics for unigram and contextual contribution regression of activation values.

We can therefore conclude that steps *Justifying Specifics* and *Accounting for Results*, as characterized by neurons 81 and 109, can reasonably be distinguished by their relative propensities for either conveying structural information about the paper or using language which implies reflections on prior data when supplemented with document level structural information via the CRF. The depth and robustness of this multifaceted approach can be leveraged to make further inferences regarding the defining characteristics of steps. These in turn can provide proposals for ways to re-characterize the data annotation approach in such a way that advances our understanding, not just of neural model structure, but human language in academic argumentation.

Neuron	Top 5 1-grams	Bottom 5 1-grams
81	methods	26%
	discussion	developments
	method	2.4.1
	results	sole
	introduction	creation
109	future	timber
	elicitation	draw
	substantiate	item
	refers	unavailable
	summarizing	soybeans

Table 6: Top and bottom 5 unigrams for predicting neuron activation.

8 CONCLUSION

In order to investigate automated approaches to analysis of rhetorical structure in student writing, we have contextualized our work in the RWT corpus, a rhetorical structure corpus for academic writing that can cultivate complex, rich, and contextual discourse analysis. We demonstrated a classification task that can be performed on the corpus and showed that a model that leverages rich contextual information integrated in academic writing can increase performance compared to contex-unaware methods. Furthermore, we show that meaningful comparisons can be made between steps by focusing on the most important set of neurons through the application of a multi-lens approach.

By understanding how the neurons operate within the model, we can use their activations when the model is run over student academic writing to detect features that we previously could not

detect in order to provide feedback we could not have previously been able to provide. For instance, feedback using this method could use the expected progression of moves and steps compared to the use of moves and steps by a student to give specific feedback to the student of, not just where to make a modification, but what type of modification should be made. This could be implemented into a essay writing platform so students can get instant feedback on their draft's argument structure. In this way we use both the power of deep learning to accurately characterize a document, and a method to unpack the characterizations to help writers improve their arguments.

This work focuses specifically on student writing at the undergraduate level. In our current work, we are exploring how the techniques proposed here apply more broadly to student writing at different stages of development. Furthermore, while this work focuses specifically on rhetorical structure, other approaches to language processing have already demonstrated success at assessment of other aspects of student writing. Going forward we plan to conduct more comparisons with previously published approaches in connection with a wider range of writing phenomena.

ACKNOWLEDGMENTS

This research was funded in part by NSF grants ACI-1443068 and IIS 1546393 and funding from the Schmidt Foundation.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644* (2016).

[3] Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. Explaining Recurrent Neural Network Predictions in Sentiment Analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. 159–168.

[4] Carol Berkenkotter and Thomas N Huckin. 2016. *Genre knowledge in disciplinary communication: Cognition/culture/power*. Routledge.

[5] Lynn Carlson, Mary Ellen Okurowski, and Daniel Marcu. 2002. *RST discourse treebank*. Linguistic Data Consortium, University of Pennsylvania.

[6] Jifan Chen, Qi Zhang, Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Implicit Discourse Relation Detection via a Deep Architecture with Gated Relevance Network.. In *ACL (1)*.

[7] François Chollet. 2015. Keras. https://github.com/fchollet/keras.

[8] Elena Cotos, Stephen Gilbert, and Jivko Sinapov. 2014. NLP-based analysis of rhetorical functions for AWE feedback.. In *Research challenges in CALL, Proceedings of the 16th International CALL Research Conference*. 117–123.

[9] Elena Cotos, Sarah Huffman, and Stephanie Link. 2015. Furthering and applying move/step constructs: Technology-driven marshalling of Swalesian genre theory for EAP pedagogy. *Journal of English for Academic Purposes* 19 (2015), 52–72.

[10] Elena Cotos, Stephanie Link, and Sarah Rebecca Huffman. 2016. Studying disciplinary corpora to teach the craft of Discussion. *Writing and Pedagogy* 87, 1 (2016), 33.

[11] Elena Cotos, Stephanie Link, and Sarah Rebecca Huffman. 2017. Effects of DDL technology on genre learning. *Language Learning & Technology* (2017).

[12] Elena Cotos and Nick Pendar. 2016. Discourse classification into rhetorical functions for AWE feedback. *calico journal* 33 (2016), 1.

[13] Greg Durrett and Dan Klein. 2015. Neural CRF Parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Vol. 1. 302–312.

[14] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18, 5 (2005), 602–610.

[15] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications* 13, 4 (1998), 18–28.

[16] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991* (2015).

[17] Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse relation language models. *arXiv preprint arXiv:1603.01913* (2016).

[18] Yangfeng Ji, Gongbo Zhang, and Jacob Eisenstein. 2015. Closing the gap: Domain adaptation from explicit to implicit discourse relations. (2015).

[19] Rekia Kadari, Yu Zhang, Weinan Zhang, and Ting Liu. 2017. CCG Supertagging via Bidirectional LSTM-CRF Neural Architecture. *Neurocomputing* (2017).

[20] Budsaba Kanoksilapatham. 2007. Rhetorical moves in biochemistry research articles. In *Discourse on the move: Using corpus analysis to describe discourse structure*, Douglas Biber, Ulla Connor, and Thomas A Upton (Eds.). Vol. 28. John Benjamins Publishing.

[21] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078* (2015).

[22] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[23] Sanjay Krishnan and Eugene Wu. 2017. PALM: Machine Learning Explanations For Iterative Debugging. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*. ACM, 4.

[24] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of NAACL-HLT*. 260–270.

[25] Jaesong Lee, Joong-Hwi Shin, and Jun-Seok Kim. 2017. Interactive Visualization and Manipulation of Attention-based Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 121–126.

[26] Wenqiang Lei, Yuanxin Xiang, Yuwei Wang, Qian Zhong, Meichun Liu, and Min-Yen Kan. 2018. Linguistic Properties Matter for Implicit Discourse Relation Recognition: Combining Semantic Interaction, Topic Continuity and Attribution. (2018).

[27] Mike Lewis, Kenton Lee, and Luke Zettlemoyer. 2016. Lstm ccg parsing. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 221–231.

[28] Kevin Lin and Eugene Wu. 2017. Searching for Meaning in RNNs using Deep Neural Inspection. (2017).

[29] Xuezhe Ma and Eduard Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 1064–1074.

[30] Danielle S McNamara, Laura K Allen, Scott A Crossley, Mihai Dascalu, and Cecile A Perret. 2017. Natural language processing and learning analytics. *Handbook of learning analytics* (2017), 93.

[31] Danielle S McNamara, Scott A Crossley, and Rod Roscoe. 2013. Natural language processing in an intelligent writing strategy tutoring system. *Behavior research methods* 45, 2 (2013), 499–515.

[32] Eleni Miltsakaki, Rashmi Prasad, Aravind K Joshi, and Bonnie L Webber. 2004. The Penn Discourse Treebank.. In *LREC*.

[33] Shotaro Misawa, Motoki Taniguchi, Yasuhide Miura, and Tomoko Ohkuma. 2017. Character-based Bidirectional LSTM-CRF with words and characters for Japanese Named Entity Recognition. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*. 97–102.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[35] Nick Pendar and Elena Cotos. 2008. Automatic identification of discourse moves in scientific article introductions. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, 62–70.

[36] 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)*. 1532–1543.

[37] Robi Polikar. 2012. Ensemble learning. In *Ensemble machine learning*. Springer, 1–34.

[38] Anna Potapenko, Artem Popov, and Konstantin Vorontsov. 2017. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. In *Conference on Artificial Intelligence and Natural Language*. Springer, 167–180.

[39] Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444* (2017).

[40] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1135–1144.

[41] Ryan M Rifkin and Ross A Lippert. 2007. Notes on regularized least squares. (2007).

[42] Carolyn Penstein Rosé. 2017. Discourse Analytics. *Handbook of Learning Analytics* (2017), 105.

[43] David E Rumelhart, James L McClelland, PDP Research Group, et al. 1987. *Parallel distributed processing*. Vol. 1. MIT press Cambridge, MA.

[44] Alexander M Rush, David Sontag, Michael Collins, and Tommi Jaakkola. 2010. On dual decomposition and linear programming relaxations for natural language processing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1–11.

[45] Attapol Rutherford and Nianwen Xue. 2014. Discovering Implicit Discourse Relations Through Brown Cluster Pair Representation and Coreference Patterns.. In *EACL*, Vol. 645. 2014.

[46] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *arXiv preprint arXiv:1708.08296* (2017).

[47] Mark D Shermis and Jill Burstein. 2013. *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

[48] John Sinclair. 2005. Corpus and Text - Basic Principles. In *Developing Linguistic Corpora: a Guide to Good Practice*, Martin Wynne (Ed.). Oxbow Books Oxford, 1–16.

[49] Wei Song, Dong Wang, Ruiji Fu, Lizhen Liu, Ting Liu, and Guoping Hu. 2017. Discourse Mode Identification in Essays. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 112–122. http://aclweb.org/anthology/P17-1011

[50] Hendrik Strobelt, Sebastian Gehrmann, Bernd Huber, Hanspeter Pfister, Alexander M Rush, et al. 2016. *Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks*. Technical Report.

[51] John Swales. 1990. *Genre analysis: English in academic and research settings*. Cambridge University Press.

[52] Sandra A Thompson and William C Mann. 1987. Rhetorical structure theory. *IPRA Papers in Pragmatics* 1, 1 (1987), 79–105.

[53] Bronwyn Woods, David Adamson, Shayne Miel, and Elijah Mayfield. 2017. Formative Essay Feedback Using Predictive Scoring Models. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. ACM, New York, NY, USA, 2071–2080. https://doi.org/10.1145/3097983.3098160

[54] Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The CoNLL-2015 Shared Task on Shallow Discourse Parsing.. In *CoNLL Shared Task*. 1–16.

[55] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J Smola, and Eduard H Hovy. 2016. Hierarchical Attention Networks for Document Classification.. In *HLT-NAACL*. 1480–1489.