Unsupervised Learning of Discourse-Aware Text Representation for Essay Scoring

Farjana Sultana Mim[1] **Naoya Inoue**[1,2] **Paul Reisert**[2,1]
Hiroki Ouchi[2,1] **Kentaro Inui**[1,2]
[1]Tohoku University [2]RIKEN Center for Advanced Intelligence Project (AIP)
{mim, naoya-i, inui} @ecei.tohoku.ac.jp
{paul.reisert, hiroki.ouchi} @riken.jp

Abstract

Existing document embedding approaches mainly focus on capturing sequences of words in documents. However, some document classification and regression tasks such as essay scoring need to consider discourse structure of documents. Although some prior approaches consider this issue and utilize discourse structure of text for document classification, these approaches are dependent on computationally expensive parsers. In this paper, we propose an unsupervised approach to capture discourse structure in terms of coherence and cohesion for document embedding that does not require any expensive parser or annotation. Extrinsic evaluation results show that the document representation obtained from our approach improves the performance of essay Organization scoring and Argument Strength scoring.

1 Introduction

Document embedding is important for many NLP tasks such as document classification (e.g., essay scoring and sentiment classification) (Le and Mikolov, 2014; Liu et al., 2017; Wu et al., 2018; Tang et al., 2015) and summarization. While embedding approaches can be supervised, semi-supervised and unsupervised, recent studies have largely focused on unsupervised and semi-supervised approaches in order to utilize large amounts of unlabeled text and avoid expensive annotation procedures.

In general, a document is a discourse where sentences are logically connected to each other to provide comprehensive meaning. Discourse has two important properties: *coherence* and *cohesion* (Halliday, 1994). Coherence refers to the semantic relatedness among sentences and logical order of concepts and meanings in a text. For example, *"I saw Jill on the street. She was going home."* is coherent whereas *"I saw Jill on the street. She has two sisters."* is incoherent. Cohesion refers to the use of linguistic devices that hold a text together. Example of these linguistic devices include conjunctions such as discourse indicators (DIs) (e.g., *"because" and "for example"*), coreference (*e.g., "he" and "they"*), substitution, ellipsis etc.

Some text classification and regression tasks need to consider discourse structure of text in addition to dependency relations and predicate-argument structures. One example of such tasks is essay scoring, where discourse structure (e.g., coherence and cohesion) plays a crucial role, especially when considering *Organization* and *Argument Strength* criteria, since they refer to logical-sequence awareness in texts. Organization refers to how good an essay structure is, where well-structured essays logically develop arguments and state positions by supporting them (Persing et al., 2010). Argument Strength means how strongly an essay argues in favor of its thesis to persuade the readers (Persing and Ng, 2015).

An example of the relation between coherence and an essay's Organization is shown in Figure 1. The high-scored essay (i.e., Organization score of 4) first states its position regarding the prompt and then provides several reasons to strengthen the claim. It is considered coherent because it follows a logical order. However, the low-scored essay is not clear on its position and what it is arguing about. Therefore, it can be considered incoherent since it lacks logical sequencing.

Previous studies on document embedding have primarily focused on capturing word similarity, word dependencies and semantic information of documents (Le and Mikolov, 2014; Liu et al., 2017; Wu et al., 2018; Tang et al., 2015). However, less attention has been paid to capturing discourse structure for document embedding in an unsupervised manner and no prior work applies unsupervised document representation learning to essay scoring. In short, it has not yet been explored how some of the discourse properties can

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 378–385
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

Prompt: Some people say that in our modern world , dominated by science, technology and industrialization, there is no longer a place for dreaming and imagination. What is your opinion?

Coherent Essay: Organization Score = 4	**Incoherent Essay:** Organization Score = 2.5
There is no doubt in the fact that we live under the full reign of science, technology and industrialization. Our lives are dominated by them in every aspect. ..... In other words, what I am trying to say more figuratively is that in our world of science, technology and industrialization there is no really place for dreaming and imagination. *One of the reasons for the disappearing of the dreams and the imagination from our life is one that I really regret to mention, that is the lack of time. We are really pressed for time nowadays .....*	*The world we are living in is without any doubt a modern and civilized one. It is not like the world five hundred years ago, it is not even like the one fifty years ago. Perhaps we - the people who live nowadays, are happier than our ancestors, but perhaps we are not.* *The strange thing is that we judge and analyse their world without knowing it and maybe without trying to know it. The only thing that is certain is that the world is changing and it is changing so fast that even we cannot notice it. Sciece has developed to such an extent that it is difficult to believe this can be true. .....*

Figure 1: Example of coherent and incoherent ICLE essays with their Organization score.

be included in text embedding without an expensive parser and how document embeddings affect essay scoring tasks.

In this paper, we propose an unsupervised method to capture discourse structure in terms of cohesion and coherence for document embedding. We train a document encoder with unlabeled data which learns to discriminate between coherent/cohesive and incoherent/incohesive documents. We then use the pre-trained document encoder to obtain feature vectors of essays for Organization and Argument Strength score prediction, where the feature vectors are mapped to scores by regression. The advantage of our approach is that it is fully unsupervised and does not require any expensive parser or annotation. Our results show that capturing discourse structure in terms of cohesion and coherence for document representation helps to improve the performance of essay Organization scoring and Argument Strength scoring. We make our implementation publicly available.[1]

2 Related Work

The focus of this study is the unsupervised encapsulation of discourse structure (coherence and cohesion) into document representation for essay scoring. A popular approach for document representation is the use of fixed-length features such as bag-of-words (BOW) and bag-of-ngrams due to their simplicity and highly competitive results (Wang and Manning, 2012). However, such approaches fail to capture the semantic similarity of words and phrases since they treat each word or phrase as a discrete token.

Several methods for document representation learning have been introduced in recent years. One popular unsupervised method is doc2vec (Le and Mikolov, 2014), where a document is mapped to a unique vector and every word in the document is also mapped to a unique vector. Then, the document vector and and word vectors are either concatenated or averaged to predict the next word in a context. Liu et al. (2017) used a convolutional neural network (CNN) to capture longer range semantic structure within a document where the learning objective predicted the next word. Wu et al. (2018) proposed Word Mover's Embedding (WME) utilizing Word Mover's Distance (WMD) that considers both word alignments and pre-trained word vectors to learn feature representation of documents. Tang et al. (2015) proposed a semi-supervised method called Predictive Text Embedding (PTE) where both labeled information and different levels of word co-occurrence were encoded in a large-scale heterogeneous text network, which was then embedded into a low dimensional space. Although these approaches have been proven useful for several document classification and regression tasks, their focus is not on capturing the discourse structure of documents.

One exception is the study by Ji and Smith (2017) who illustrated the role of discourse structure for document representation by implementing a discourse structure (defined by RST) aware model and showed that their model improves text categorization performance (e.g., sentiment classification of movies and Yelp reviews, and prediction of news article frames). The authors utilized an RST-parser to obtain the discourse dependency

[1] Our implementation is publicly available at https://github.com/FarjanaSultanaMim/DiscoShuffle

tree of a document and then built a recursive neural network on top of it. The issue with their approach is that texts need to be parsed by an RST parser which is computationally expensive. Furthermore, the performance of RST parsing is dependent on the genre of documents (Ji and Smith, 2017).

Previous studies have modeled text coherence (Li and Jurafsky, 2016; Joty et al., 2018; Mesgar and Strube, 2018). Farag et al. (2018) demonstrated that state-of-the-art neural automated essay scoring (AES) is not well-suited for capturing adversarial input of grammatically correct but incoherent sequences of sentences. Therefore, they developed a neural local coherence model and jointly trained it with a state-of-the-art AES model to build an adversarially robust AES system. Mesgar and Strube (2018) used a local coherence model to assess essay scoring performance on a dataset of holistic scores where it is unclear which criteria of the essay the score considers.

We target Organization and Argument Strength dimension of essays which are related to coherence and cohesion. Persing et al. (2010) proposed heuristic rules utilizing various DIs, words and phrases to capture the organizational structure of texts. Persing and Ng (2015) used several features such as part-of-speech, n-grams, semantic frames, coreference, and argument components for calculating Argument Strength in essays. Wachsmuth et al. (2016) achieved state-of-the-art performance on Organization and Argument Strength scoring of essays by utilizing argumentative features such as sequence of argumentative discourse units (e.g., *(conclusion, premise, conclusion)*). However, Wachsmuth et al. (2016) used an expensive argument parser to obtain such units.

3 Base Model

3.1 Overview

Our base model consists of (i) a base document encoder, (ii) auxiliary encoders, and (iii) a scoring function. The base document encoder produces a vector representation $\mathbf{h}^{\text{base}}$ by capturing a sequence of words in each essay. The auxiliary encoders capture additional essay-related information that is useful for essay scoring and produce a vector representation $\mathbf{h}^{\text{aux}}$. By taking $\mathbf{h}^{\text{base}}$ and $\mathbf{h}^{\text{aux}}$ as input, the scoring function outputs a score.

Specifically, these encoders first produce the representations, $\mathbf{h}^{\text{base}}$ and $\mathbf{h}^{\text{aux}}$. Then, these representations are concatenated into one vector, which is mapped to a feature vector $\mathbf{z}$.

$$\mathbf{z} = \tanh(\mathbf{W} \cdot [\mathbf{h}^{\text{base}}; \mathbf{h}^{\text{aux}}]) \ , \qquad (1)$$

where $\mathbf{W}$ is a weight matrix. Finally, $\mathbf{z}$ is mapped to a scalar value by the sigmoid function.

$$y = \text{sigmoid}(\mathbf{w} \cdot \mathbf{z} + b) \ ,$$

where $\mathbf{w}$ is a weight vector, b is a bias value, and y is a score in the range of $(0, 1)$. In the following subsections, we describe the details of each encoder.

3.2 Base Document Encoder

The base document encoder produces a document representation $\mathbf{h}^{\text{base}}$ in Equation 1. For the base document encoder, we use the Neural Essay Assessor (NEA) model proposed by Taghipour and Ng (2016). This model uses three types of layers: an embedding layer, a Bi-directional Long Short-Term Memory (BiLSTM) (Schuster and Paliwal, 1997) layer and a mean-over-time layer.

Given the input essay of T words $w_{1:T} = (w_1, w_2, \cdots, w_T)$, the embedding layer (Emb) produces a sequence of word embeddings $\mathbf{w}_{1:T} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_T)$.

$$\mathbf{w}_{1:T} = \text{Emb}(w_{1:T}) \ ,$$

where each word embedding is a d^{word} dimensional vector, i.e. $\mathbf{w}_i \in \mathbb{R}^{d^{\text{word}}}$.

Then, taking $\mathbf{x}_{1:T}$ as input, the BiLSTM layer produces a sequence of contextual representations $\mathbf{h}_{1:T} = (\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_T)$.

$$\mathbf{h}_{1:T} = \text{BiLSTM}(\mathbf{x}_{1:T}) \ ,$$

where each representation $\mathbf{h}_i$ is $\mathbb{R}^{d^{\text{hidden}}}$.

Finally, taking $\mathbf{h}_{1:T}$ as input, the mean-over-time layer produces a vector averaged over the sequence.

$$\mathbf{h}^{\text{mean}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{h}_t \ . \qquad (2)$$

We use this resulting vector as the base document representation, i.e. $\mathbf{h}^{\text{base}} = \mathbf{h}^{\text{mean}}$.

3.3 Auxiliary Encoders

The auxiliary encoders produce a representation of essay-related information $\mathbf{h}^{\text{aux}}$ in Equation 1. We provide two encoders that capture different types of essay-related information.

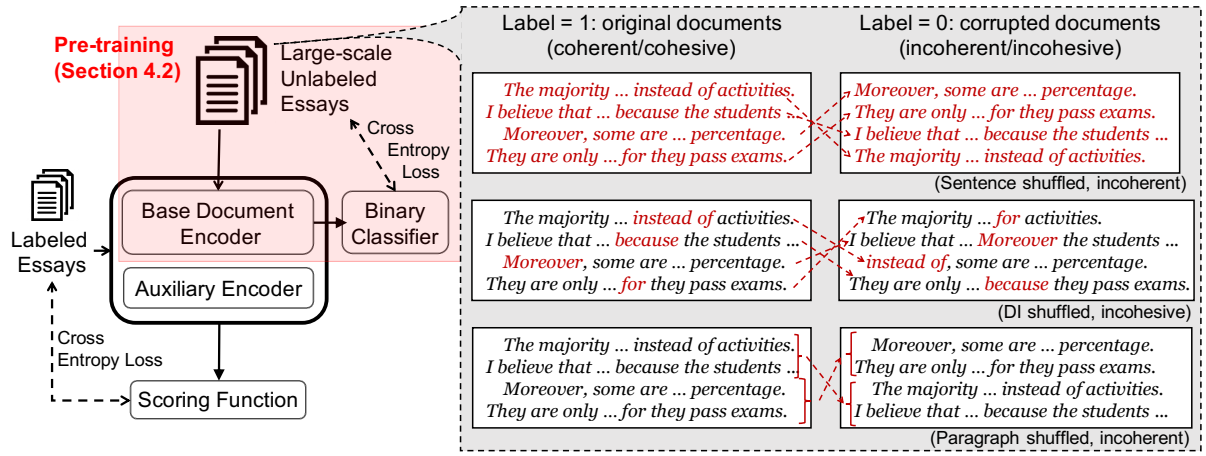

Figure 2: Proposed method for unsupervised learning of discourse-aware text representation utilizing coherent/incoherent and cohesive/incohesive texts and use of the discourse-aware text embeddings for essay scoring.

Paragraph Function Encoder (PFE). Each paragraph in an essay plays a different role. For instance, the first paragraph tends to introduce the topic of the essay, and the last paragraph tends to sum up the whole content and make some conclusions. Here, we capture such paragraph functions.

Specifically, we obtain paragraph function labels of essays using Persing et al. (2010)'s heuristic rules.[2] Persing et al. (2010) specified four paragraph function labels: Introduction (**I**), Body (**B**), Rebuttal (**R**) and Conclusion (**C**). We represent these labels as vectors and incorporate them into the base model. The paragraph function label encoder consists of two modules, an embedding layer and a BiLSTM layer.

We assume that an essay consists of M paragraphs, and the i-th paragraph has already been assigned a function label p_i. Given the sequence of paragraph function labels of an essay $p_{1:M} = (p_1, p_2, ..., p_M)$, the embedding layer ($\text{Emb}^{\text{para}}$) produces a sequence of label embeddings, i.e. $\boldsymbol{p}_{1:M} = \text{Emb}^{\text{para}}(p_{1:M})$, where each embedding $\boldsymbol{p}_i$ is $\mathbb{R}^{d^{\text{para}}}$. Then, taking $\boldsymbol{p}_{1:M}$ as input, the BiLSTM layer produces a sequence of contextual representations $\mathbf{h}_{1:M} = \text{BiLSTM}(\boldsymbol{p}_{1:M})$, where $\mathbf{h}_i$ is $\mathbb{R}^{d^{\text{PFE}}}$. We use the last hidden state $\mathbf{h}_M$ as the paragraph function label sequence representation, i.e. $\mathbf{h}^{\text{aux}} = \mathbf{h}_M$.

Prompt Encoder (PE). As shown in Figure 1, essays are written for a given prompt, where the prompt itself can be useful for essay scoring. Based on this intuition, we incorporate prompt information.

The prompt encoder uses an embedding layer and a Long Short-Term Memory (LSTM) (Hochreiter, Sepp and Schmidhuber, Jürgen, 1997) layer to produce a prompt representation. Formally, we assume that the input is a prompt of N words, $w_{1:N} = (w_1, w_2, \cdots, w_N)$. First, the embedding layer maps the input prompt $w_{1:N}$ to a sequence of word embeddings, $\mathbf{w}_{1:N}$, where $\mathbf{w}_i$ is $\mathbb{R}^{d^{\text{prompt}}}$. Then, taking $\mathbf{w}_{1:N}$ as input, the LSTM layer produces a sequence of hidden states, $\mathbf{h}_{1:N} = (\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_N)$, where $\mathbf{h}_i$ is $\mathbb{R}^{d^{\text{PE}}}$. The last hidden state is regarded as the resulting representation, i.e. $\mathbf{h}^{\text{aux}} = \mathbf{h}_N$.

4 Proposed Method

4.1 Overview

Figure 2 summarizes the proposed method. First, we pre-train a base document encoder (Section 3.2) in an unsupervised manner. The pretraining is motivated by the following hypotheses: (i) artificially corrupted incoherent/incohesive documents lack logical sequencing, and (ii) training a base document encoder to differentiate between the original and incoherent/incohesive documents makes the encoder logical sequence-aware.

The pre-training is done in two steps. First, we pre-train the document encoder with large-scale unlabeled essays. Second, we pre-train the encoder using only the unlabeled essays of target corpus used for essay scoring. We expect that this fine-tuning alleviates the domain mismatch between the large-scale essays and target essays

[2] See http://www.hlt.utdallas.edu/~persingq/ICLE/orgDataset.html for further details.

(e.g., essay length). Finally, the pre-trained encoder is then re-trained on the annotations of essay scoring tasks in a supervised manner.

4.2 Pre-training

We artificially create incoherent/incohesive documents by corrupting them with random shuffling methods: (i) *sentences*, (ii) *only DIs* and (iii) *paragraphs*. Figure 2 shows examples of original and corrupted documents. We shuffle DIs since they are important for representing the logical connection between sentences. For example, *"Mary did well although she was ill"* is logically connected, but *"Mary did well but she was ill."* and *"Mary did well. She was ill."* lack logical sequencing because of improper and lack of DI usage, respectively. Paragraph shuffling is also important since coherent essays have sequences like *Introduction-Body-Conclusion* to provide a logically consistent meaning of the text.

Specifically, we treat the pre-training as a binary classification task where the encoder classifies documents as coherent/cohesive or not.

$$P(y(d) = 1|d) = \sigma(\mathbf{w}^{\text{unsup}} \cdot \mathbf{h}^{\text{mean}}) \ ,$$

where y is a binary function mapping from a document d to $\{0, 1\}$, in which 1 represents the document is coherent/cohesive and 0 represents not. The base document representation $\mathbf{h}^{\text{mean}}$ (Eq. 2) is multiplied with a weight vector $\mathbf{w}^{\text{unsup}}$, and the sigmoid function σ returns a probability that the given document d is coherent/cohesive.

To train the model parameters, we minimize the binary cross-entropy loss function,

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(P(y(d_i) = 1|d_i)) + (1 - y_i)\log(1 - P(y(d_i) = 1|d_i)) \ ,$$

where y_i is a gold-standard label of coherence/cohesion of d_i and N is the total number of documents. Note that y_i is automatically assigned in the corruption process where an original document has a label of 1 and an artificially corrupted document has a label of 0.

5 Experiments

5.1 Setup

We use five-fold cross-validation for evaluating our models with the same split as Persing et al. (2010); Persing and Ng (2015) and Wachsmuth et al. (2016). The reported results are averaged over five folds. However, our results are not directly comparable since our training data is smaller as we reserve a development set (100 essays) for model selection while they do not. We use the mean squared error as an evaluation measure.

Data We use the International Corpus of Learner English (ICLE) (Granger et al., 2009) for essay scoring which contains 6,085 essays and 3.7 million words. Most of the ICLE essays (91%) are argumentative and vary in length, having 7.6 paragraphs and 33.8 sentences on average (Wachsmuth et al., 2016). Some essays have been annotated with different criteria among which 1,003 essays are annotated with Organization scores and 1,000 essays are annotated with Argument Strength scores. Both scores range from 1 to 4 at half-point increments. For our scoring task, we utilize the 1,003 essays.

To pre-train the document encoder, we use 35,222 essays from four datasets, (i) the Kaggle's Automated Student Assessment Prize (ASAP) dataset[3] (12,976) (ii) TOEFL11 (Blanchard et al., 2013) dataset (12,100), (iii) The International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa, 2013) dataset (5,600), and (iv) the ICLE essays not used for Organization and Argument Strength scoring (4,546).[4]

See Appendix A and B for further details on the hyperparameters and preprocessing.

5.2 Results and Discussion

From two baseline models, we report the best model for each task (*Base+PFE* for Organization, *Base+PE* for Argument Strength).

Table 1 indicates that the proposed unsupervised pre-training improves the performance of Organization and Argument Strength scoring. These results support our hypothesis that training with random corruption of documents helps a document encoder learn logical sequence-aware text representations. In most cases, fine-tuning the encoder for each scoring task again helps to improve the performance.

The results indicate that paragraph shuffling

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

[4] During pre-training with paragraph shuffled essays, we use only 16,646 essays (TOEFL11 and ICLE essays) since ASAP and ICNALE essays have a single paragraph.

Model	Shuffle Type	Fine-tuning	Mean Squared Error	
			Organization	Argument Strength
Baseline	-	-	0.182	0.248
Proposed	Sentence		0.187	**0.244**
	Sentence	✓	0.186	**0.244***
	Discourse Indicator		0.187	**0.242**
	Discourse Indicator	✓	0.193	**0.246**
	Paragraph		**0.172***	**0.236***
	Paragraph	✓	**0.169***	**0.231***
Persing et al. (2010)			0.175	-
Persing et al. (2015)			-	0.244
Wachsmuth et al. (2016)			0.164	0.226

Table 1: Performance of essay scoring. '*' indicates a statistical significance (Wilcoxon signed-rank test, $p < 0.05$) against the baseline model. Base+PFE and Base+PE are used in Organization and Argument Strength, respectively.

is the most effective in both scoring tasks (statistically significant by Wilcoxon's signed rank test, $p < 0.05$). This could be attributed to the fact that paragraph sequences create a more clear organizational and argumentative structure. Suppose that an essay first introduces a topic, states their position, supports their position and then concludes. Then, the structure of the essay would be regarded as "well-organized". Moreover, the argument of the essay would be considered "strong" since it provides support for their position. The results suggest that such levels of abstractions (e.g., *Introduction-Body-Body-Conclusion*) are well captured at a paragraph-level, but not at a sentence-level or DI-level alone.

Furthermore, a manual inspection of DIs identified by the system suggest room for improvement in DI shuffling. First, the identification of DIs is not always reliable. Almost half of DIs identified by our simple pattern matching algorithm (see Appendix B) were not actually DIs (e.g., *we have survived **so** far only external difficulties*). Second, we also found that some DI-shuffled documents are sometimes cohesive. This happens when original document counterparts have two or more DIs with the more or less same meaning (e.g., *since* and *because*). We speculate that this confuses the document encoder in the pre-training process.

6 Conclusion and Future Work

We proposed an unsupervised strategy to capture discourse structure (i.e., coherence and cohesion) for document embedding. We train a document encoder with coherent/cohesive and randomly corrupted incoherent/incohesive documents to make it logical-sequence aware. Our method does not require any expensive annotation or parser. The experimental results show that the proposed learning strategy improves the performance of essay Organization and Argument Strength scoring.

Our future work includes adding more unannotated data for pre-training and trying other unsupervised objectives such as swapping clauses before and after DIs (e.g., A because B $\rightarrow$ B because A). We also intend to perform intrinsic evaluation of the learned document embedding space. Moreover, we plan to evaluate the effectiveness of our approach on more document regression or classification tasks.

7 Acknowledgements

This work was supported by JST CREST Grant Number JPMJCR1513 and JSPS KAKENHI Grant Number 19K20332. We would like to thank the anonymous ACL reviewers for their insightful comments. We also thank Ekaterina Kochmar for her profound and useful feedback.

References

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.

Youmna Farag, Helen Yannakoudakis, and Ted Briscoe. 2018. Neural automated essay scoring and coherence modeling for adversarially crafted input. *arXiv preprint arXiv:1804.06898*.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. International corpus of learner English.

Miochael AK Halliday. 1994. An introduction to functional grammar 2nd edition. *London: Arnold*.

Hochreiter, Sepp and Schmidhuber, Jürgen. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

S Ishikawa. 2013. ICNALE: the international corpus network of Asian learners of English. *Retrieved on November*, 21:2014.

Yangfeng Ji and Noah Smith. 2017. Neural discourse structure for text categorization. *arXiv preprint arXiv:1702.01829*.

Shafiq Joty, Muhammad Tasnim Mohiuddin, and Dat Tien Nguyen. 2018. Coherence modeling of asynchronous conversations: A neural entity grid approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 558–568.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Jiwei Li and Dan Jurafsky. 2016. Neural net models for open-domain discourse coherence. *arXiv preprint arXiv:1606.01545*.

Chundi Liu, Shunan Zhao, and Maksims Volkovs. 2017. Unsupervised Document Embedding With CNNs. *arXiv preprint arXiv:1711.04168*.

Mohsen Mesgar and Michael Strube. 2018. A Neural Local Coherence Model for Text Quality Assessment. In *Proceedings of the 2018 Conference on EMNLP*, pages 4328–4339.

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling organization in student essays. In *Proceedings of the 2010 Conference on EMNLP*, pages 229–239. ACL.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the ACL the 7th International Joint Conference on Natural Language Processing*, volume 1, pages 543–552.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on EMNLP*, pages 1882–1891.

Jian Tang, Meng Qu, and Qiaozhu Mei. 2015. Pte: Predictive text embedding through large-scale heterogeneous text networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1165–1174. ACM.

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691.

Sida Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the ACL: Short Papers-Volume 2*, pages 90–94. Association for Computational Linguistics.

Lingfei Wu, Ian EH Yen, Kun Xu, Fangli Xu, Avinash Balakrishnan, Pin-Yu Chen, Pradeep Ravikumar, and Michael J Witbrock. 2018. Word Mover's Embedding: From Word2Vec to Document Embedding. *arXiv preprint arXiv:1811.01713*.

Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on EMNLP*, pages 1393–1398.

A Hyperparameters

We use BiLSTM with 200 hidden units in each layer for the base document encoder ($d^{\text{hidden}} = 200$). For the paragraph function encoder, we use a BiLSTM with hidden units of 200 in each layer ($d^{\text{PFE}} = 200$). For the prompt encoder, an LSTM with an output dimension of 300 is used ($d^{\text{PE}} = 300$). We use the 50-dimensional pre-trained word embeddings released by Zou et al. (2013) in our base document encoder ($d^{\text{word}} = 50, d^{\text{prompt}} = 50$).

We use the Adam optimizer with a learning rate of 0.001 and a batch size of 32. We use early stopping with patience 15 (5 for pre-training), and train the network for 100 epochs. The vocabulary consists of the 90,000 and 15,000 most frequent words for pre-training and essay scoring, respectively. Out-of-vocabulary words are mapped to special tokens. We perform hyperparameter tuning and choose the best model. We tuned norm clipping maximum values (3,5,7) and dropout rates (0.3, 0.5, 0.7, 0.9) for all models on the development set.

B Preprocessing

We lowercase the tokens and specify an essay's paragraph boundaries with special tokens. During sentence/DI shuffling for pre-training, paragraph boundaries are not used. We collect 847 DIs from

the Web.[5] We exclude the DI "and" since it is not always used for initiating logic (e.g milk, banana *and* tea). In essay scoring data, we found 176 DIs and average DIs per essay is around 24. In the pre-training data, the number of DIs found is 204 and the average DIs per essay is around 13. We identified DIs by simple string-pattern matching.

[5] http://www.studygs.net/wrtstr6.htm, http://home.ku.edu.tr/~doregan/Writing/Cohesion.html etc.