Unsupervised Learning of Discourse-Aware Text Representation

Farjana Sultana Mim[†] Naoya Inoue[†,‡] Paul Reisert[‡] Hiroki Ouchi[‡] Kentaro Inui[†] [‡]

[†]Tohoku University [‡]RIKEN Center for Advanced Intelligence Project (AIP)
{mim,naoya-i,inui}@ecei.tohoku.ac.jp
{paul.reisert,hiroki.ouchi}@riken.jp

1 Introduction

Text or document representation is important for many NLP tasks such as document classification (e.g. essay scoring, sentiment classification), summarization, etc. Document representation approaches can be supervised, semi-supervised or unsupervised. Recent studies largely focused on unsupervised [5, 6, 15] or semi-supervised [11] methods since one can take advantage of 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*. Cohesion captures linguistic devices that link sentences into a text. Examples include *conjunction (Discourse indicators: "in my opinion", "for example", "however", "because" etc.)*, *coreference (he, she, they etc.)*, *substitution*, *ellipsis*, etc. ("Figure 1"). Coherence refers to semantic relatedness among sentences. Coherence means the reader can make sense from the entire text since it follows some kind of logical order or sequence of concepts and meanings. For example, *"I saw Jerin on the street. She was going home."* is a coherent text whereas *"I saw Jerin on the street. She has one brother and two sisters."* is incoherent.

Some text classification or regression tasks (e.g. essay scoring) need to consider discourse structure of text in addition to semantic structure. Because for these tasks, like scoring essays along particular dimension (e.g. Organization or Argument strength scoring of essays), discourse structure (coherence, cohesion etc.) play crucial role. Organization of an essay refers to its structure, where a well-structured essay logically develops arguments and states positions by supporting them [8]. Argument strength means how strongly an essay makes arguments for its thesis to convince the readers [9]. Now, discourse properties such as cohesion and coherence refer to the logical-sequence aware texts which is the basis for Organization and Argument strength scoring. "Figure 2" shows examples where Organization score of an essay is related to the discourse property coherence.

Previous document representation studies primarily focused on capturing word similarity, word dependencies or semantic features of documents [5, 6, 15, 11] which has been proven to be useful for several document classification or regression tasks (e.g. information retrieval, sentiment classification). However, none of the prior studies captured discourse structure of documents in terms of logical sequencing (coherence or cohesion). One exception is [4], who utilized discourse structure of documents defined by Rhetorical Structure Theory (RST) for classification of documents. The issue in their approach is, texts need to be parsed by an RST parser which is computationally expensive. Besides, the performance of RST parsing is dependent on the genre of documents [4]. In sum, it has not yet been explored how some of the discourse properties can be included in text representation without using any expensive parser.

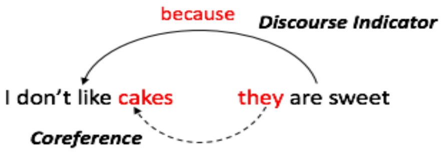
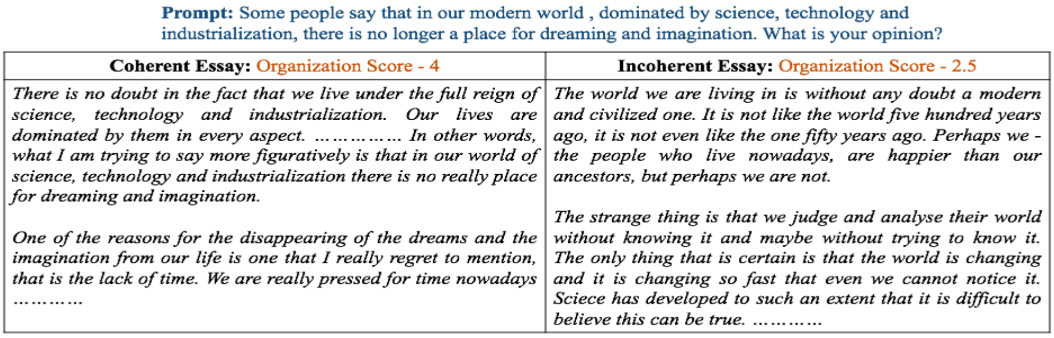


Figure 1: Some discourse properties of the sentence *"I don't like cakes because they are sweet"*.

In this paper, we propose an unsupervised method to capture discourse structure in terms of logical sequence of sentences (cohesion & coherence) for document representation. We train a document encoder with unlabeled data which learns to discriminate between coherent and incoherent documents. Then, we use the pre-trained encoder to obtain feature vectors of documents in order to perform extrinsic evaluation, i.e. the task of Organization scoring and Argument Strength scoring of essays where these 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 extrinsic evaluation results show that capturing discourse structure in terms of logical sequencing (coherence & cohesion) for document representation helps to improve the performance of essay Organization scoring and Argument Strength scoring.

2 Related work

Although the issue of document representation was addressed by several previous studies, the most common and popular fixed-length feature is still bag-of-words (BOW) or bag-of-ngrams due to its simplicity and highly competitive results [13]. However, BOW approaches fail to capture the semantic similarity of words and phrases since it treats each word or phrase as a discrete token. Hence it provides sparse repre-



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 2: Example essays from ICLE corpus with their Organization score

sentation with large dimensionality. In recent years, several unsupervised approaches for document representation have been introduced. One of the popular unsupervised methods is doc2vec [5] where a document vector is incorporated along with the word vectors to learn the vector representation of the document. The training objective was predicting the words in the document. [6] used a CNN to capture longer range semantic structure within a document where the learning objective was also predicting the next word. Their model learned a joint semantic space by optimizing both word vectors and embeddings. In [15], word alignments and pretrained word vectors were utilized to learn semantic features. [11] 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. However, all these approaches were basically to produce a semantic feature representation of documents.

[4] 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. They used an RST-parser to get the discourse dependency tree of a document and then built a recursive neural network on top of it. Nevertheless, RST parser is domain dependent [4] and computationally inconvenient.

In [7], a local coherence model was used to assess essay scoring performance but the dataset had holistic scores. The issue with holistic scores is that it is unclear which dimension of the essay (argument, content) the score refers to. [8] proposed some heuristic rules which is based on various discourse indicators, words and phrases to capture organization structure of text. In [9], several features like POS n-grams, semantic frames, coreference, argument component were used to anticipate how strong an essays arguments are. [12] showed that argumentative features such as sequence of argumentative discourse units(ADUs) (e.g *(conclusion, premise, conclusion)* or *(None, Thesis)*) improve the performance of Organization and Argument strength scoring. They also used an expensive argument parser to obtain the ADUs.

3 Data

We use the International Corpus of Learner English (ICLE) which contains 6,085 essays and 3.7 million words [2]. Most of the ICLE essays (91%) are argumentative. ICLE essays vary in length, having 7.6 paragraphs and 33.8 sentences on average [12]. Some of its essays have been annotated with different scores among which 1,003 essays are annotated with Organization scores and 1,000 essays are annotated with Argument strength scores. Both Organization and Argument strength scores ranges from 1 to 4 at half-point increments. We use these 1,003 essays for scoring task and the rest of the ICLE essays (4578) are used to pretrain the document encoder.

4 Base document encoder

Our main goal is to demonstrate if capturing logical-sequencing for text representation in an unsupervised way helps with document classification or regression tasks, not to propose the best model to capture it. Therefore, our intention is to implement a simple encoder.

4.1 NEA

We replicate the Neural Essay Assessor (NEA) model proposed by [10] as our document encoder. The model has three layers. First, an embedding layer provides word embeddings for a given essay. Then, we use a Bi-directional gated recurrent unit (GRU)[1] to incorporate contextual information of words by summing up information from both directions of words. Finally, a mean-over-time layer is used to get the mean of the intermediate GRU layers. During essay scoring, we apply a linear layer to map the vector produced by mean-over-time layer to a score. We use a sigmoid function to get scores in the range of (0, 1).

4.2 NEA+PN10

We create our second baseline model by obtaining paragraph sequences of essays using Persing's [8] heuristic rules. [8] specified four paragraph function labels: Introduction (I), Body (B), Rebuttal (R) and Conclusion (C) and used some heuristic rules to identify them (see the original paper for details). We use the same heuristic rules to get paragraph sequences of essays. Given a paragraph sequence of an essay, the embedding layer first produces a vector for each paragraph function label. Then, a Long Short-Term Mem-

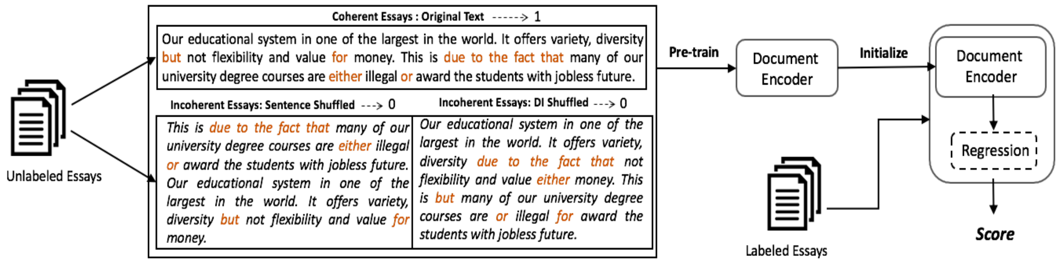

Coherent Essays : Original Text ------> 1
Our educational system in one of the largest in the world. It offers variety, diversity but not flexibility and value for money. This is due to the fact that many of our university degree courses are either illegal or award the students with jobless future.
Incoherent Essays: Sentence Shuffled ---> 0
This is due to the fact that many of our university degree courses are either illegal or award the students with jobless future. Our educational system in one of the largest in the world. It offers variety, diversity but not flexibility and value for money.
Incoherent Essays: DI Shuffled ---> 0
Our educational system in one of the largest in the world. It offers variety, diversity due to the fact that not flexibility and value either money. This is but many of our university degree courses are or illegal for award the students with jobless future.
Unlabeled Essays
Pre-train
Document Encoder
Initialize
Document Encoder
Regression
Labeled Essays
Score

Figure 3: Proposed method for unsupervised learning of logical sequence-aware text representation utilizing coherent and incoherent texts and using the logical sequence-aware text representation for essay scoring.

ory (LSTM) [3] is used to encode them. We take the last hidden state of the LSTM and concatenate it with the vector representation of essay obtained from our first baseline NEA. Once again, at the time of essay scoring, we use a linear layer with a sigmoid function. We name this baseline *NEA+PN10*.

5 Proposed Method

5.1 Key Idea

Our idea is first to train a document encoder in an unsupervised way with coherent and incoherent documents so that the encoder learns to differentiate between them. Our hypothesis is that artificially corrupted incoherent documents lack logical sequencing and training a document encoder to distinguish original documents from artificially corrupted ones makes it logical sequence-aware. We then use this pretrained encoder for the essay scoring task. In the essay scoring task setting, we initialize the essay scoring encoder with the pretrained logical sequence-aware encoder and obtain a vector representation of documents. Afterwards, we perform regression to obtain scores from corresponding document vectors (see “Figure 3”).

Shuffle Type	Accuracy
Sentences	0.733
Discourse Indicators	0.880

Table 1: Performance of coherence discrimination task.

5.2 Unsupervised Pretraining

We pretrain the document encoder in an unsupervised manner to get logical sequence-aware document representations. For this purpose, we artificially create incoherent texts so that the encoder can learn to discriminate coherent documents from incoherent ones. We obtain incoherent documents by corrupting documents. For corruption, we consider two approaches: (i) randomly shuffling *sentences* and (ii) randomly shuffling *only discourse indicators*. “Figure 3” shows some examples of coherent and incoherent (corrupted) documents. We give importance to discourse indicators as they are important for representing the logical connection between sentences. For example, *“Mary did well in the exam although she was sick”* is logically connected while *“Mary did well in the exam but she was sick.”* and *“Mary did well in the exam. She was sick.”* lack logical sequencing because of improper use and no use of discourse indicators, respectively.

Finally, we treat the pretraining as a binary classification task where the encoder classifies documents as coherent or incoherent. The performance of this coherence discrimination task is shown in “Table 1”. We use our baseline model NEA as pretrained document encoder.

6 Experiments

6.1 Preprocessing

We use the same preprocessing step for both pretraining and essay scoring data. We remove all punctuation, lowercase the tokens, and normalize the gold-standard scores to the range of [0, 1]. We specify the sentence boundaries and paragraph boundaries of essays with special tokens. During our testing phase, we re-scale the predicted normalized scores to the original range of scores and then measure the performance.

6.2 Discourse Indicators

We collect, in total, 847 discourse indicators from the Web. We exclude the discourse indicator ”and” since it’s used most frequently and not always for initiating logic (e.g milk, banana *and* tea).

6.3 Setting

We use pretrained word embeddings for our baseline models which were released by [16]. The embedding dimension is set to 50. For the pretrained encoder, we initialize the word embeddings randomly and learn them alongside the model parameters.

We use Adam optimizer with the learning rate set to 0.001. In our experiments, we set the batch size to 32, include early stopping with patience 15, and train the network for 100 epochs. The vocabulary is the 40,000 and 15,000 most frequent words for pretraining and essay scoring, respectively. All other words are mapped to special tokens. We use norm clipping technique and dropout for all systems. Norm clipping maximum values (3,5,7,10) and dropout rates (0.5,0.7, 0.75, 0.9) are set to different values for different systems. We use GRU with hidden states dimension 300 for both pretraining and essay scoring. For encoding paragraph sequences, an LSTM with an output dimension of 400 is used.

6.4 Evaluation

We use five-fold cross-validation for evaluating our models with the same split as Persing 2010 [8, 9] and Wachsmuth 2016 [12]. However, our results cannot be directly compared with Persing 2010 & Wachsmuth 2016 since our training data is smaller (-100 essays)] as

Model	Pretraining	Shuffle Type	Fine-tuning	MSE (Org.)	MSE (Arg. Strength)
NEA		-	-	0.348	0.255
NEA+PN10		-	-	0.200	-
NEA	✓	Sentence		0.344	**0.249***
NEA	✓	Sentence	✓	0.347	**0.251**
NEA+PN10	✓	Sentence		**0.191***	-
NEA+PN10	✓	Sentence	✓	**0.187***	-
NEA	✓	Discourse Indicator		0.363	0.255
NEA	✓	Discourse Indicator	✓	0.365	**0.252**
NEA+PN10	✓	Discourse Indicator		0.203	-
NEA+PN10	✓	Discourse Indicator	✓	**0.197**	-

Table 2: Performance of essay scoring. * indicates a statistically significant improvement (Wilcoxon's signed-rank test (p <0.05))

we reserve development set for model selection, while they do not. Wachsmuth 2016 use expensive argumentation parser, while we do not.

For Organization and Argument strength scoring task, we measure the mean squared error (MSE) of regression. As we mentioned above, we use several hyper-parameters for our system. While we tuned the hyper-parameters for the baseline models, we didn't tune it for our proposed models. For tuning the hyper-parameters of the baseline models, we randomly select one fold (in our case fold 1) and choose hyper-parameters based on that fold.

6.5 Results and discussion

Table 2 lists MSE (averaged over five folds) of two baseline models and our proposed systems (pretrained) for Organization and Argument strength scoring task. From the results in Table 2, we see that unsupervised pretraining with coherent and incoherent documents improves Organization and Argument strength scoring performance. These results support our hypothesis that training with random corruption of documents helps learning logical sequence-aware text representation. While shuffling the discourse indicators does not make much difference, shuffling sentences noticeably improves scoring performance. By shuffling sentences, we get statistically significant improvement (by Wilcoxon's signed rank test ([14]), p <0.05) for both Organization and Argument strength scoring. Re-tuning the encoder for Organization scoring again helps to improve the performance.

7 Conclusion and Future Work

We proposed an unsupervised strategy to capture discourse structure for document representation in terms of logical sequence of sentences (i.e. coherence and cohesion). Our method does not require any expensive annotation or parser. We train a document encoder with coherent and incoherent documents to make it logical-sequence aware. Then, we use the logical-sequence aware encoder to obtain document vectors for the task of essay scoring. Our results show that learning logical sequence-aware document representation in an unsupervised way improves essay Organization and Argument strength scoring performance.

Our future work includes tuning hyper-parameters for our proposed models, adding more unannotated data for pretraining and trying other unsupervised objectives, e.g., shuffling (paragraph-based shuffling, cohesive device shuffling), swapping clauses before and after discourse indicators (e.g. A because B =>B because A). Also, we intend to incorporate prompt information into our baseline models (similar to [8]). Moreover, we plan to try more document regression or classification tasks and see how these unsupervised objectives affect the performance.

Acknowledgement

This work was partially supported by JST CREST Grant Number JPMJCR1513, including AIP challenge.

References

[1] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473.*

[2] Granger, S., Dagneaux, E., Meunier, F., and Paquot, M. (2009). International corpus of learner english.

[3] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[4] Ji, Y. and Smith, N. (2017). Neural discourse structure for text categorization. *arXiv preprint arXiv:1702.01829.*

[5] Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

[6] Liu, C., Zhao, S., and Volkovs, M. (2017). Unsupervised document embedding with cnns. *arXiv preprint arXiv:1711.04168.*

[7] Mesgar, M. and Strube, M. (2018). A neural local coherence model for text quality assessment. In *Proceedings of the 2018 Conference on EMNLP*, pages 4328–4339.

[8] Persing, I., Davis, A., and Ng, V. (2010). Modeling organization in student essays. In *Proceedings of the 2010 Conference on EMNLP*, pages 229–239. ACL.

[9] Persing, I. and Ng, V. (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.

[10] Taghipour, K. and Ng, H. T. (2016). A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on EMNLP*, pages 1882–1891.

[11] Tang, J., Qu, M., and Mei, Q. (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.

[12] Wachsmuth, H., Al Khatib, K., and Stein, B. (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.

[13] Wang, S. and Manning, C. D. (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.

[14] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83.

[15] Wu, L., Yen, I. E., Xu, K., Xu, F., Balakrishnan, A., Chen, P.-Y., Ravikumar, P., and Witbrock, M. J. (2018). Word mover's embedding: From word2vec to document embedding. *arXiv preprint arXiv:1811.01713.*

[16] Zou, W. Y., Socher, R., Cer, D., and Manning, C. D. (2013). Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on EMNLP*, pages 1393–1398.