Automated Essay Scoring with Discourse-Aware Neural Models

Farah Nadeem[1], Huy Nguyen[2], Yang Liu[2], and Mari Ostendorf[1]

[1]Department of Electrical and Computer Engineering, University of Washington
{farahn, ostendor}@uw.edu
[2]LAIX Inc.
{huy.nguyen, yang.liu}@liulishuo.com

Abstract

Automated essay scoring systems typically rely on hand-crafted features to predict essay quality, but such systems are limited by the cost of feature engineering. Neural networks offer an alternative to feature engineering, but they typically require more annotated data. This paper explores network structures, contextualized embeddings and pre-training strategies aimed at capturing discourse characteristics of essays. Experiments on three essay scoring tasks show benefits from all three strategies in different combinations, with simpler architectures being more effective when less training data is available.

1 Introduction

In the context of large scale testing and online learning systems, automated essay scoring (AES) is an important problem. There has been work on both improving the performance of these systems and on validity studies (Shermis, 2014). The ability to evaluate student writing has always been important for language teaching and learning; now it also extends to science, since the focus is shifting towards assessments that can more accurately gauge construct knowledge as compared to multiple choice questions (Shermis, 2014). Most existing systems for automatic essay scoring leverage hand crafted features, ranging from word-counts to argumentation structure and coherence, in linear regression and logistic regression models (Chodorow and Burstein, 2004; Shermis and Burstein, 2013; Klebanov et al., 2016; Nguyen and Litman, 2018). Improving feature-based models requires extensive redesigning of features (Taghipour and Ng, 2016). Due to high variability in types of student essays, feature-based systems are often individually designed for specific prompts (Burstein et al., 2013). This poses a challenge for building essay scoring systems.

These problems (and the success of deep learning in other areas of language processing) have led to the development of neural methods for automatic essay scoring, moving away from feature engineering. A variety of studies (mostly LSTM-based) have reported AES performance comparable to or better than feature-based models (Taghipour and Ng, 2016; Cummins and Rei, 2018; Wang et al., 2018; Jin et al., 2018; Farag et al., 2018; Zhang and Litman, 2018). However, the current state-of-the-art models still use a combination of neural models and hand-crafted features (Liu et al., 2019).

While vanilla RNNs, particularly LSTMs, are good at representing text sequences, essays are longer structured documents and less well suited to an RNN representation. Thus, our work looks at advancing AES by exploring other architectures that incorporate document structure for longer documents. Discourse structure and coherence are important aspects of essay writing and are often explicitly a part of grading rubrics. We explore methods that aim at discourse-aware models, through design of the model structure, use of discourse-based auxiliary pretraining tasks, and use of contextualized embeddings trained with cross-sentence context (Devlin et al., 2018). In order to better understand the relative advantages of these methods, we compare performance on three essay scoring tasks with different characteristics, contrasting results with a strong feature-based system.

Our work makes two main contributions. First, we demonstrate that both discourse-aware structures and discourse-related pre-training (via auxiliary tasks or contextualized embeddings) benefit performance of neural network systems. In a TOEFL essay scoring task, we obtain a substantial improvement over the state-of-the-art. Second, we show that complex contextualized embedding

Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 484–493
Florence, Italy, August 2, 2019.

models are not useful for tasks with small annotated training sets. Simpler discourse-aware neural models are still useful, but they benefit from combination with a feature-based model.

2 Method

2.1 Neural Models

The overall system involves a neural network to map an essay to a vector, which is then used with ordinal regression (McCullagh, 1980) for essay scoring. For this work we consider two neural models that incorporate document structure:

- Hierarchical recurrent network with attention (HAN) (Yang et al., 2016)
- Bidirectional context with attention (BCA) (Nadeem and Ostendorf, 2018)

Both models are LSTM based. HAN captures the hierarchical structure within a document, by using two stacked layers of LSTMs. The first layer takes word embeddings as input and outputs contextualized word representations. Self attention is used to compute a sentence vector as a weighted average of the contextualized word vectors. The second LSTM takes sentence vectors as input and outputs a document vector based on averaging using self attention at the sentence level.

BCA extends HAN to account for cross sentence dependencies. For each word, using the contextualized word vectors output from the first LSTM, a look-back and look-ahead context vector is computed based on the similarity with words in the previous and following sentence, respectively. The final word representation is then created as a concatenation of the LSTM output, the look-back and look-ahead context vectors, and then used to create a sentence vector using attention weights, which feeds into the second LSTM. The representation of cross-sentence dependencies makes this model discourse aware.

2.2 Auxiliary Training Tasks

Neural networks typically require more training data than feature-based models, but unlike these models, neural networks can make use of related tasks to improve performance through pretraining. We use additional data chosen with the idea that having related tasks for pretraining can help the model learn aspects that impact the main classification problem. We use the following tasks:

- Natural language inference (NLI): given a pair of sentences, predict their relation as neutral, contradictory, or entailment.
- Discourse marker prediction (DM): given a pair of sentences, predict the category of discourse marker that connects them, e.g. "however" (corresponding to the idea opposition category).

The NLI task has been shown to improve performance for several NLP tasks (Cozma et al., 2018). The DM prediction task is used since discourse structure is an important aspect for essay writing. Both tasks involve sentence pairs, so they impact the first-level LSTM of the HAN and BCA models.

The use of contextualized embeddings can also be thought of as pre-training with an auxiliary task of language modeling (or masked language modeling). In this work, we chose the bidirectional transformer architecture (BERT) embeddings (Devlin et al., 2018), which uses a transformer architecture trained on two tasks, masked language model and next sentence prediction. We hypothesized that the next sentence prediction would capture aspects of discourse coherence.

2.3 Training Methods

All HAN models and a subset of BCA models are initialized with pretrained Glove word embeddings[1] (Pennington et al., 2014). All models are trained with the essay training data.

For models that are pretrained, the word-level LSTM and bidirectional context with attention (for BCA), are common for all tasks used in training. Given the word-level representations, the model computes attention weights over words for the target task (DM, NLI or essay scoring). The sentence representation is then computed by averaging the word representations using task-specific attention weights. For the pretraining tasks, the sentence representations the two sentences in the pair are concatenated, passed through a feedforward neural network, and used with task-specific weights and biases to predict the label. For pretraining the BCA with the auxiliary tasks, the forward context vector is computed for the first sentence and the backward context vector is computed for the second sentence. This allows the model to learn the similarity projection matrix during pretraining.

[1] http://nlp.stanford.edu/data/glove.42B.300d.zip

For the essay scoring task there is another sentence-level LSTM on top of the word-level LSTM, with sentence-level attention, followed by task-specific weights and biases. Pretraining is followed by training with the essay data, with all model parameters updated during training, except for the auxiliary task-specific word-level attention, feedforward networks, weights and biases. The network used for BCA with pretraining tasks is shown in Figure 1. The hyper-parameters were tuned for the auxiliary tasks and the essay scoring task. To incorporate BERT embeddings in our model, we freeze the BERT model, and learn contextualized token embeddings for our data using the base uncased model. The tokens are from the second-to-last hidden layer, since we are not fine-tuning the model and the last layer is likely to be more tuned to the original BERT training tasks. These embeddings are then used as input to the BCA model (BERT-BCA), which is then trained on the essay scoring task.

3 Experiments

3.1 Data

The first set of essay data is the ETS Corpus of Non-Native Written English from the Linguistic Data Consortium (LDC) (Blanchard et al., 2013) consisting of 12,100 TOEFL essays.[2]

The data has essay scores given as high, medium or low. Two train/test splits are used:

- Split 1 from LDC, 11,000 training essays and 1100 test essays
- Split 2 from (Klebanov et al., 2016), 6074 training essays and 2023 test essays

Split 1 is a larger publicly available set, and split 2 is used in the prior published work on this data. The data distribution is shown in Table 1. The data is skewed, with the medium score being the majority class.

To evaluate model performance on smaller data sets, we use essays in Sets 1 and 2 of the Automated Student Assessment Prize (ASAP) Competition.[3] We chose the first two sets from the ASAP data, since they are persuasive essays, and are likely to benefit more from discourse-aware pretraining. The two essay sets have topics in computer usage and library censorship, respectively. Data statistics of the two essay sets are shown in Table 2. Since only the training samples are available for both sets, we report results for 5-fold cross-validation using the same splits as (Taghipour and Ng, 2016).

Data set	Essays	High	Medium	Low
Train/dev	11,000	3,835	5,964	1,202
Test	1,100	367	604	129
Train/dev	6,074	2,102	3,318	655
Test	2,023	700	1,101	222

Table 1: Label distribution in LDC TOEFL dataset. Data is split into training and test sets: split 1 (upper part) and split 2 (lower part).

Data set	Essays	Avg. len	Score range
1	1783	350	2-12
2	1800	350	1-6

Table 2: Data statistics for essay sets 1 and 2 of ASAP corpus.

Pretraining tasks use two data sets. The NLI task uses the Stanford natural language inference (SNLI) data set (Bowman et al., 2015). We cast our NLI task as a four-way classification task, because a subset of the data does not have gold labels. Unlabeled examples were used with an "X" label. While tuning on the main task, we found that including the fourth NLI label gave better performance on the essay scoring than not using it.

The DM task is based on a collection of over 13K free books from `www.smashwords.com` – an online book distribution platform.[4] Labeled discourse marker data was created by identifying sentence pairs that had a discourse marker at the start of the second sentence. We used 87 discourse markers, which were then mapped to seven groups, for a total of 581,650 sentence pairs. A set of randomly-selected 95,450 consecutive sentence pairs without discourse markers was added to the data set as negative examples, leading to an eight way classification task. Example discourse marker categories include:

- **Idea opposition**: nonetheless, on the other hand, however
- **Idea justification**: in other words, for example, alternatively
- **Time relation**: meanwhile, in the past, simultaneously

[2] https://catalog.ldc.upenn.edu/LDC2014T06
[3] http://www.kaggle.com/c/asap-aes
[4] The data set published by (Zhu et al., 2015) is no longer available, so we compiled our own data set.

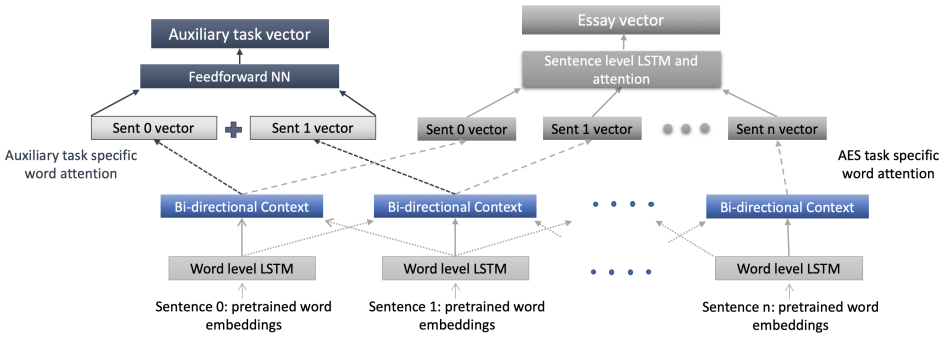


Figure 1: Network structure for BCA with pretraining tasks.

The complete set of labels, number of samples and mapping scheme are given in Appendix A.

3.2 Training Configurations

We explore the following setups to train AES models for the LDC-TOEFL essays:

1. Training using only LDC essay data;
2. Pretraining with one task (either NLI or DM prediction), followed by training with the essay data;
3. Pretraining alternating between the two auxiliary tasks (NLI-DM), followed by training with the essay data; and
4. Training the BCA model with only the essay data, using static BERT token embeddings as input to the model.

For the ASAP data, we used the third training configuration.

For the pretraining tasks, 10% of the training data is used as a held out development set. On pretraining tasks, the BCA model achieves accuracy 0.60 (8 classes) on the development set of DM data, and accuracy 0.78 (4 classes) on the dedicated test set of SNLI data (Bowman et al., 2015).

Ten-fold cross validation was used for the LDC essay data, five-fold for the ASAP data. A vocabulary size of 75000 was used for all the experiments, except those trained with BERT token embeddings. Dropout and early stopping was used for regularization, including variational recurrent dropout (Gal and Ghahramani, 2016) at both LSTM layers. Hyper-parameter training was used to find the optimal dropout and determine early stopping. Network sizes, dropout and number of epochs over the training data are listed in Table 3.[5]

[5]Trained models and code is available at https://github.com/Farahn/AES

Shared parameters	
Word level LSTM	150
Word level attention weight size	75
Sentence level LSTM	150
Sentence level attention weight size	50
Dropout rate	0.25-0.5
BERT embedding size	768
Auxiliary task parameters	
Feed-forward network layer 1	500
Feed-forward network layer 2	250
Training epochs	
Essay data	35-45
NLI data	15-25
DM data	5-7

Table 3: Hyper-parameters

3.3 Baselines

We develop a feature-based model that combines text readability (Vajjala and Meurers, 2014; Vajjala, 2018) and argument mining features (Nguyen and Litman, 2018). In our implementation, we remove one set of basic features, e.g., word counts, spelling errors etc., since they are present in both models and keep the set from (Vajjala and Meurers, 2014). Given the extracted features, a gradient boosting algorithm is used to learn a regression model. Predicted scores are scaled and rounded to calculate Quadratic Weighted Kappa (QWK) against the true scores. These two feature sets are chosen because they incorporate discourse features in AES. In (Vajjala and Meurers, 2014), the authors used the addDiscourse toolkit (Pitler et al., 2009), which takes as input the syntactic tree of the sentence, and tags the discourse connectives, e.g., therefore, however, and their senses, e.g., CONTINGENCY.Cause, COMPARISON.Contrast. These automated annotations are then used to calculate connective based features,

e.g., number of discourse connectives per sentence, number of each sense. In (Nguyen and Litman, 2018), an end-to-end pipeline system was built to parse input essays for argument structures. The system identifies argument components, e.g., claims, premises, in essay sentences, and determines if a support relation is present between each pair of components. Based on that, the authors extract 33 argumentative features used for their AES model.

In addition, we build neural baselines using existing sentence representations as input to a document level LSTM. Specifically, we compare: i) the *BERT sentence encoder*, taking the sentence representation from the second-to-last hidden layer of BERT (as in BERT-BCA) and ii) the *Universal sentence encoder* (USE) (Cer et al., 2018), which is trained on multiple down-stream tasks including classification and sentiment analysis. Unlike for BERT, there are no sequential sentence tasks used in training USE, so we claim that USE is not discourse-aware. The vectors output from the LSTM are then averaged using attention weights to generate a document representation, as in the HAN and BCA models, so these baselines are also hierarchical models and will be referred to as *BERT-HAN* and *USE-HAN*, respectively. For both setups, the sentence vectors are frozen and not updated during training; initial experiments found no performance gain from fine-tuning.

3.4 Results

3.4.1 LDC TOEFL Essays

The results are shown in Table 4, together with previously reported results for feature-based automatic essay scoring systems from (Klebanov et al., 2016) (Klebanov16) and (Nguyen and Litman, 2018) (Nguyen18). Significance testing was done on the test set using bootstrap.

All neural models outperform previously reported results on split 2, with the exception of USE-HAN, as does the augmented feature-based baseline implemented here. Using the new feature-based system as the baseline for significance testing, only the results from BERT-BCA give a statistically significant improvement ($p < 0.01$). The two models that do not explicitly use discourse cues, HAN and USE-HAN, have the lowest scores of the neural models. The best result is obtained when we combine contextualized token level embeddings from BERT with the cross-sentence attention in BCA. This indicates that the two methods are complementary and useful for writing evaluation.

Model	Split 1	Split 2
Arg (Klebanov16)	-	0.344
Length (Klebanov16)	-	0.518
Arg + Len (Klebanov16)	-	0.540
Nguyen18	-	0.622
Feature baseline	0.659	0.642
USE-HAN	0.626	0.618
BERT-HAN	0.688	0.680
HAN	0.635	0.623
NLI-HAN	0.643	0.630
DM-HAN	0.651	0.654
NLI-DM-HAN	0.655	0.644
BCA	0.637	0.636
NLI-BCA	0.652	0.647
DM-BCA	0.661	0.661
NLI-DM-BCA	0.659	0.663
BERT-BCA	0.729	0.715

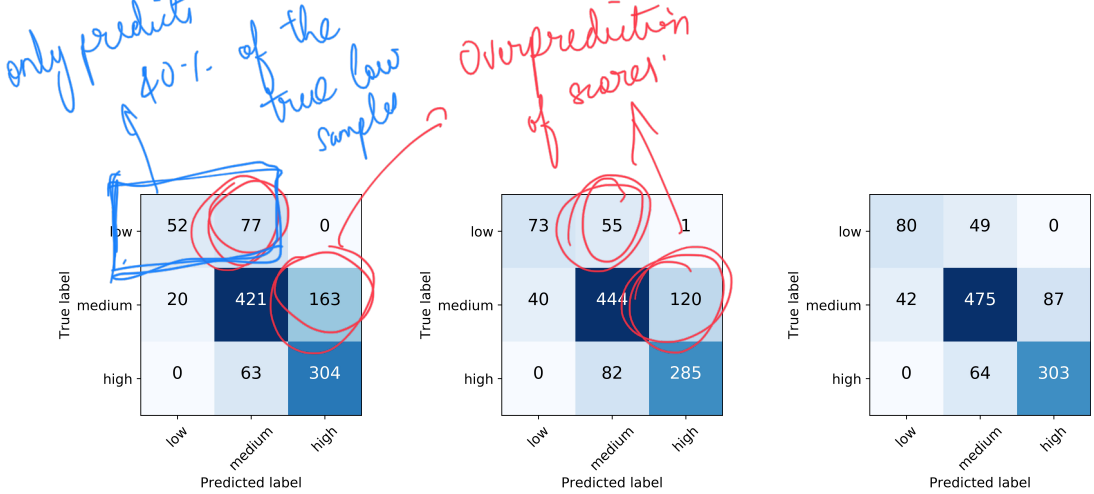
Table 4: Results for the essay scoring task on LDC TOEFL corpus for both splits reported in QWK.

Figure 2 shows the confusion matrices for the USE-HAN baseline, DM-BCA and BERT-BCA systems for the LDC TOEFL split 1. The confusion matrices indicate that both USE-HAN and DM-BCA over-predict the essay scores compared to BERT-BCA, i.e. assign a higher scoring category than the true score. The problem is most severe for USE-HAN, which correctly labels only 40% of the low test samples.

3.4.2 ASAP Essays

Results are reported for 5-fold CV. For each of the splits, 20% data is used to tune the dropout rate, learning rate and number of iterations. Since there was a small variation in the optimal parameters for the 5 folds, we used the average of the parameters from the first two sets for training all five folds. The test QWK is computed by taking the true labels and predictions for all 5 test sets. For the ASAP data set, we report performances of our feature baseline, the best sentence representation model, and the best pretrained BCA model. In addition, we present a simple combination of the feature-based and BCA model, averaging the scores predicted by the two models. The results are shown in table 5.

For both ASAP sets, feature based models perform better than the neural models. We hypoth-



(a) USE-HAN Baseline (b) DM-BCA (c) BERT-BCA

Figure 2: Confusion matrices for the USE-HAN baseline vs. the best neural models on LDC TOEFL split 1.

Model	ASAP 1	ASAP 2
TSLF (Liu 2019)	0.852	0.736
Feature baseline	0.833	0.692
BERT-HAN	0.748	0.627
NLI-DM-BCA	0.800	0.671
NLI-DM-BCA+features	0.840	0.711

Table 5: Results for the essay scoring task for ASAP sets 1 and 2 reported in QWK.

Model	sim_2		sim_{all}	
	Min	σ	Min	σ
USE-HAN	-0.180	0.214	-0.440	0.400
BERT-HAN	-0.012	0.013	-0.047	-0.005
HAN	0.021	-0.023	0.069	-0.051
DM-HAN	-0.414	0.365	-0.437	0.460
BCA	-0.394	0.362	**-0.571**	**0.632**
DM-BCA	**-0.448**	**0.433**	-0.554	0.589
BERT-BCA	0.052	-0.071	0.186	-0.153

Table 6: Correlation of sim_2 and sim_{all} with the true essay scores for LDC TOEFL split 1.

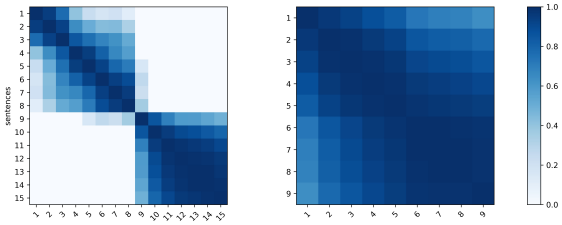


Figure 3: Sentence similarity for DM-BCA, left is a high scoring essay (ID 108264), right is a low scoring essay (ID 10226).

esize that this is due to having less training data than for the TOEFL essays. Using the pretrained BERT-HAN model does significantly worse than the pretrained NLI-DM-BCA model. Combining the best neural and feature-based model gives a small, but insignificant performance gain. A more sophisticated combination would likely yield better results.

The current state-of-the art is the two stage learning framework (TSLF) (Liu et al., 2019). The model has two components, one using sentence representation from BERT input to an RNN (similar to our BERT-HAN), and the second component uses hand crafted features. The BERT sentence representations are used to learn an essay score, a prompt-relevance score and a "coherence" score, trained on original and permuted essays. Document representations from the neural network and the hand crafted features are then used together in a gradient-boosting decision tree to predict the final essay score.

4 Analysis and Discussion

We hypothesized that good quality essays would be more coherent. To see if this is captured by the learned sentence representations, we examined sentence similarities in the TOEFL essays in relation to the essay score. Taking the sentence vector outputs from the second LSTM layer for essay i for a particular model for LDC split 1, we compute the cosine similarity of each sentence with its neighboring sentence sim_2 and with all other sentences sim_{all}. We then compute the correlation of the mean, min and standard deviation of both sim_2 and sim_{all} with the true labels. The mean gave no meaningful differences between models, but there were differences for the min and standard deviation (σ), which are presented in Table 6.

In terms of correlation between essay scores and min/variance of sentence similarity, the highest correlations are associated with the models that use explicit discourse-aware approaches: DM pre-training and/or the BCA architecture (with-

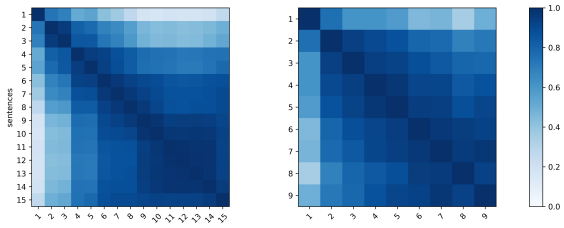


Figure 4: Sentence similarity for BERT-HAN, left is a high scoring essay (ID 108264), right is a low scoring essay (ID 10226).

out BERT). The correlation values indicate that these sentence representations capture aspects of text structure that are reflected in a positive trend for the variance and negative trends for minimum sentence similarity. This suggests that discussion on multiple topics/aspects, as opposed to a single theme, tends to result in high scoring essays, as visualized in Figure 3 for DM-BCA. The fact that low-scoring essays have higher cross-sentence similarity likely reflects a less varied use of vocabulary than higher coherence.

Both BERT-HAN and BERT-BCA lead to representations for which sentence similarity has lower variance and lower correlation of the standard deviation with essay quality. The BERT-BCA sentence embedding similarities, illustrated for the same essays in Figure 4, seem to be learning a fundamentally different representation, but clearly also useful. In both cases, the BERT embeddings are learned using the next sentence prediction objective (together with the masked language model objective). We hypothesize that AES performance improvement with BERT, i.e., BERT-HAN and BERT-BCA, may be due to contextualized word representations (within and cross-sentence), reducing the need for BCA cross-sentence attention, as seen by the good performance of the BERT-HAN model, which has no explicit cross-sentence dependencies.

An initial investigation of sentence-level attention weights suggests that weights tend to be more uniform for low scoring essays and show more variation for higher scoring ones. However we observe no meaningful difference between the different models.

For both BERT-HAN and BERT-BCA, we froze the sentence and token embeddings (respectively) for use in our models. Our experiments indicated that it is hard to fine-tune the BERT model with the limited training data available for the LDC TOEFL and ASAP training sets. Experiments showed that freezing the model and using tokens as input to the model gave similar performance as fine-tuning BERT, and was much easier to optimize. For the ASAP data, initial experiments using BERT token embeddings as input to BCA gave significantly worse performance than the best BCA model. Fine tuning in this case also proved more challenging, and results indicated that it did not perform better than freezing sentence embeddings.

5 Related Work

Neural networks have already shown promising results for AES. Our work differs from prior efforts primarily in the particular architecture that we use. Most prior work uses LSTMs (Farag et al., 2018; Wang et al., 2018; Cummins and Rei, 2018) or a combination LSTMs and CNNs (Taghipour and Ng, 2016; Zhang and Litman, 2018), cast as linear or logistic regression problems. In contrast, we use a hierarchically structured model with ordinal regression. The work by (Farag et al., 2018) is similar in that they model local text coherence, though the coherence features are for detecting adversarial examples and not used directly in essay scoring. The neural essay scoring system presented in (Cummins and Rei, 2018) also uses a multitask framework, but the auxiliary task is grammatical error detection. In our work, we found that adding grammatical error features improved an existing feature-based system, and we expect that grammar error detection would be a useful auxiliary task for our neural model as well.

There is no single data set that all systems report on, which makes it difficult to compare results. For the TOEFL data, where prior published work uses feature-based systems (Klebanov et al., 2016; Nguyen and Litman, 2018), our system provides state-of-the-art results. For the ASAP data, where there are published studies using neural networks, the best scoring systems use ensembling and/or combine neural and feature-based approaches (Liu et al., 2019; Taghipour and Ng, 2016). Such methods would likely also benefit our model, but the focus here was on the use of auxiliary pretraining tasks.

Our study explored the hierarchical attention network (HAN) (Yang et al., 2016) and bidirectional context with attention (BCA) network (Nadeem and Ostendorf, 2018). Other neural network architectures for document classification

could also be explored, e.g., (Le and Mikolov, 2014; Ji and Smith, 2017; Card et al., 2018). Numerous previous studies have looked at using external data to improve performance of neural classifiers. One study that influenced our work is (Jernite et al., 2017), which showed that discourse-based tasks such as sentence order and conjunction prediction can improve neural sentence representations for several NLP tasks. This study used the Book Corpus data (Zhu et al., 2015) and the Gutenberg data (Stroube, 2003) for discourse-based tasks. Our task is similar, but we use a larger set of discourse markers.

Representations from pretrained models including (Devlin et al., 2018; Cer et al., 2018; Peters et al., 2018) have led to performance improvements across a variety of downstream NLP tasks. As shown in the previous section, token and sentence embeddings from BERT (Devlin et al., 2018) were useful for the essay scoring task, for which more data was available. In contrast to our work, which did not find the BERT sentence embeddings as useful for the ASAP data (when used in a hierarchical document model), BERT was found to be useful for ASAP in (Liu et al., 2019), where neural and hand-crafted features are used jointly in classification. While we experimented with both freezing and fine-tuning BERT, we observed no difference in model performance with fine-tuning. Work by (Peters et al., 2019) has shown that fine tuning BERT vs. freezing can give significant performance improvements for textual similarity tasks, but it is not significant for natural language inference tasks.

6 Conclusions

In this work we show that using a neural model with cross-sentence dependencies and having a discourse-based training task can improve performance on automatic essay scoring over both the feature-based state-of-the-art models and hierarchical LSTMs for the LDC TOEFL essay data. The natural language inference task, although useful for other text classification tasks, does not contribute as much to essay scoring. Using pretrained BERT tokens can further improve performance on the TOEFL data, indicating that other discourse-aware tasks, such as next sentence prediction, help essay scoring. For the ASAP data sets, our augmented feature-based system outperforms our best neural models, which may be due to the small amount of training data. The better results in (Liu et al., 2019) are achieved with a model that learns the combination of hand-crafted features and the neural document representation. Thus, for tasks with limited labeled data, there is still a place for hand-crafted features.

Like other neural models, our approach suffers from a lack of interpretability. While our analysis of sentence similarity with the DM-BCA model provides some useful insights into differences between high and low scoring TOEFL essays, the best scoring model did not have the same behavior. This remains an open problem.

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.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Jill Burstein, Joel Tetreault, and Nitin Madnani. 2013. The e-rater automated essay scoring system. *Handbook of automated essay evaluation: Current applications and new directions*, pages 55–67.

Dallas Card, Chenhao Tan, and Noah A Smith. 2018. Neural models for documents with metadata. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2031–2040.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Martin Chodorow and Jill Burstein. 2004. Beyond essay length: evaluating e-rater®'s performance on toefl® essays. *ETS Research Report Series*, 2004(1):i–38.

Mădălina Cozma, Andrei M Butnaru, and Radu Tudor Ionescu. 2018. Automated essay scoring with string kernels and word embeddings. *arXiv preprint arXiv:1804.07954*.

Ronan Cummins and Marek Rei. 2018. Neural multitask learning in automated assessment. *arXiv preprint arXiv:1801.06830*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems*, pages 1019–1027.

Yacine Jernite, Samuel R Bowman, and David Sontag. 2017. Discourse-based objectives for fast unsupervised sentence representation learning. *arXiv preprint arXiv:1705.00557*.

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

Cancan Jin, Ben He, Kai Hui, and Le Sun. 2018. Tdnn: a two-stage deep neural network for prompt-independent automated essay scoring. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1088–1097.

Beata Beigman Klebanov, Christian Stab, Jill Burstein, Yi Song, Binod Gyawali, and Iryna Gurevych. 2016. Argumentation: Content, structure, and relationship with essay quality. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 70–75.

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

Jiawei Liu, Yang Xu, and Lingzhe Zhao. 2019. Automated essay scoring based on two-stage learning. *arXiv preprint arXiv:1901.07744*.

Peter McCullagh. 1980. Regression models for ordinal data. *Journal of the royal statistical society. Series B (Methodological)*, pages 109–142.

Farah Nadeem and Mari Ostendorf. 2018. Estimating linguistic complexity for science texts. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 45–55.

Huy V Nguyen and Diane J Litman. 2018. Argument mining for improving the automated scoring of persuasive essays. In *Proc. AAAI*, pages 5892–5899.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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

Matthew Peters, Sebastian Ruder, and Noah A Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. *arXiv preprint arXiv:1903.05987*.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 683–691. Association for Computational Linguistics.

Mark D. Shermis. 2014. State-of-the-art automated essay scoring: Competition, results, and future directions from a united states demonstration. *Assessing Writing*, 20:53 – 76.

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

Bryan Stroube. 2003. Literary freedom: Project gutenberg. *Crossroads*, 10(1):3–3.

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

Sowmya Vajjala. 2018. Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*, 28(1):79–105.

Sowmya Vajjala and Detmar Meurers. 2014. Readability assessment for text simplification: From analysing documents to identifying sentential simplifications. *ITL-International Journal of Applied Linguistics*, 165(2):194–222.

Yucheng Wang, Zhongyu Wei, Yaqian Zhou, and Xuanjing Huang. 2018. Automatic essay scoring incorporating rating schema via reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 791–797.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

Haoran Zhang and Diane Litman. 2018. Co-attention based neural network for source-dependent essay scoring. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 399–409.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Discourse marker data

The data for discourse marker prediction task was created using over 13,000 books from www.smashwords.com. Sentence pairs with 87 discourse markers were selected, mapped to seven groups. The distribution of labels is shown in Table 7.

The mapping of labels to groups is given below:

- **Idea justification**: in other words, in particular, this means that, in fact, for example, alternatively, for instance, to exemplify, specifically, instead, indeed, as an example, as an alternative, actually, as an illustration, as a matter of fact
- **Time relation**: meanwhile, in the past, simultaneously, thereafter, after a while, by then, in turn, in the future, at the same time, previously, in the meantime
- **Idea support**: for this reason, therefore, thus, consequently, hence, as a consequence, as a result, that is the reason why, the reason is that, accordingly, this shows that, for that reason, thereby, one of the main reasons
- **Idea opposition**: nonetheless, on the other hand, however, conversely, on the contrary, in comparison, by contrast, in opposition, in contrast, still, by comparison, nevertheless
- **Idea expansion**: in like manner, likewise, in addition, also, moreover, equally important, what is more, additionally, in the same way, furthermore, besides, in addition to this, similarly
- **Alternative**: else, otherwise
- **Conclusion**: ultimately, in the end, in closing, finally, in brief, last but not least, in sum, to summarize, lastly, at the end of the day, in short, after all, in conclusion, to conclude, overall, eventually, at last, all in all, on the whole, briefly, in summary

Category	Number of samples
Idea justification	144022
Time relation	24600
Idea support	67223
Idea opposition	181949
Idea expansion	67800
Alternative	7203
Conclusion	88853
Negative samples	95450

Table 7: Categories and data distribution for the discourse marker prediction task.