Hierarchical Multi-task Learning for Organization Evaluation of Argumentative Student Essays

Wei Song[1] , **Ziyao Song**[1] , **Lizhen Liu**[1] and **Ruiji Fu**[2,3]
[1]College of Information Engineering and Academy for Multidisciplinary Studies,
Capital Normal University, Beijing, China
[2]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, China
[3]iFLYTEK AI Research (Hebei), Langfang, China
{wsong, zysong, liz_liu7480}@cnu.edu.cn, rjfu@iflytek.com

Abstract

Organization evaluation is an important dimension of automated essay scoring. This paper focuses on discourse element (i.e., functions of sentences and paragraphs) based organization evaluation. Existing approaches mostly separate discourse element identification and organization evaluation. In contrast, we propose a neural hierarchical multi-task learning approach for jointly optimizing sentence and paragraph level discourse element identification and organization evaluation. We represent the organization as a grid to simulate the visual layout of an essay and integrate discourse elements at multiple linguistic levels. Experimental results show that the multi-task learning based organization evaluation can achieve significant improvements compared with existing work and pipeline baselines. Multiple level discourse element identification also benefits from multi-task learning through mutual enhancement.

1 Introduction

Automated essay scoring (AES) has been developed for years as an educational application of natural language processing (NLP) [Page, 1966], which aims to reduce the burden on teachers and improve the educational equity. Instead of only giving a holistic score, recent research starts to evaluate particular dimensions of essay writing.

This paper focuses on evaluating the organization of argumentative student essays. Organization is an important aspect of writing. An essay could not live up to its potential without a good organization. A well organized essay should have a clear structure to accurately and logically develop ideas.

The challenges for organization evaluation include how to represent the organization of an essay and how to connect the representation to the organization quality. One representative solution is based on identifying and utilizing discourse elements [Attali and Burstein, 2006; Persing *et al.*, 2010]. Discourse elements indicate sentence functions (e.g., *prompt*, *thesis*, *main idea* and *support*) and paragraph functions (e.g., *introduction*, *body* and *conclusion*). [Attali and Burstein, 2006] compared the actual discourse elements of an essay with a standard five-paragraph structure, while [Persing *et al.*, 2010] explored the correlation between discourse element sequences and organization quality. Discourse elements also improve the interpretability of organization evaluation.

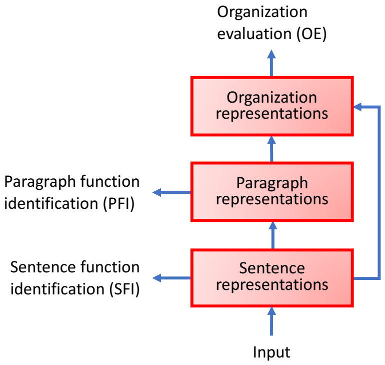
Organization evaluation (OE)
Organization representations
Paragraph function identification (PFI)
Paragraph representations
Sentence function identification (SFI)
Sentence representations
Input

Figure 1: The hierarchy of the representations and tasks.

We follow the idea of using discourse elements to represent and evaluate organization. However, the previous approaches need to be improved. First, previous discourse element identification mainly depends on heuristic rules [Persing *et al.*, 2010] or manual features [Burstein *et al.*, 2003]. The rules are too coarse and hard to be adapted to other languages, while manual features heavily rely on expert knowledge. Second, discourse element identification and organization evaluation are usually conducted separately in a pipeline. The interactions between tasks are ignored and the follow-up tasks may suffer from error propagation from previous tasks. Third, how to effectively integrate multiple level discourse elements to best indicate the organization quality is needed to be studied.

This paper proposes a novel hierarchical neural multi-task learning approach to jointly model discourse element identification and organization evaluation in an end-to-end manner. As illustrated in Figure 1, the three tasks to be optimized synchronously are *sentence function identification (SFI)*, *paragraph function identification (PFI)* and *organization evaluation (OE)*. The model receives text content as the input and automatically learns representations at different linguistic levels. The representations of higher levels depend on the representations of lower levels. Moreover, the model exploits hierarchical supervisions from sentence level to paragraph level and discourse level. The lower level tasks would provide information for higher level tasks. For example, SFI would directly affect PFI, while SFI and PFI together would help OE by providing structure information at different granularities. We also propose a grid representation of organization, which considers the visual layouts of essays, naturally integrates paragraph and sentence functions and shows to be more effective than commonly used sequence representations.

We built a dataset of more than 1,200 argumentative student essays with sentence functions, paragraph functions and organization grades annotated. We conducted comprehensive experiments to study how different levels of supervisions affect the performance of different tasks and found that organization evaluation benefits much from jointly learning with auxiliary tasks, while sentence and paragraph function identification benefit each other most. Based on multi-task learning, organization evaluation achieves significant improvements compared with existing work and pipeline models; sentence and paragraph function identification outperform existing work and single task models.

In summary, the key contributions of this paper are:

- We propose a hierarchical neural multi-task learning approach for organization evaluation with sentence function and paragraph function identification as auxiliary tasks. A grid representation of organization is proposed to integrate supervisions and representations from multiple level discourse elements.
- Experiments on a new manually annotated dataset investigate the influence of multiple tasks at different levels and demonstrate that our multi-task learning approach obtains superior performance on organization evaluation and multiple level discourse element identification compared with baselines.

2 Related Work

A primary challenge of organization evaluation is how to represent organization. A text could be represented as a sequence of topic segments through topic segmentation [Hearst, 1997] or a rhetorical discourse tree according to the rhetorical structure theory [Mann and Thompson, 1988]. These representations can reflect general text structures, but do not directly indicate organization quality, which often relates to the genre and the writing purpose. The modes of discourse have been studied and used to model local text structure [Smith, 2003] and [Song *et al.*, 2017] showed that the modes of discourse is useful for scoring narrative essays.

For argumentative texts, representing organization with discourse elements is one of the most representative solution [Burstein *et al.*, 2003; Persing *et al.*, 2010]. Discourse elements are the functions of sentences or paragraphs. [Attali and Burstein, 2006] proposed a simple approach to measure organization quality by comparing an actual essay with a basic five-paragraph structure. [Higgins *et al.*, 2004] evaluated multiple aspects of coherence between specific discourse elements. [Persing *et al.*, 2010] built the first corpus and proposed a computational model for organization evaluation. Another line of related work is to parse the argumentation structure [Stab and Gurevych, 2017a] for evaluating argumentation quality [Wachsmuth *et al.*, 2016; Stab and Gurevych, 2017b; Ke *et al.*, 2018].

The proposed approach in this paper is based on discourse elements. Previous related methods mostly adopt a pipeline manner: first identify discourse elements and then derive features for organization evaluation. Discourse elements could be identified by rules [Persing *et al.*, 2010] or be viewed as a classification problem [Burstein *et al.*, 2003] or sequence labeling problem [Song *et al.*, 2015].

In contrast to previous work, we propose a multi-task learning approach to jointly model sentence level and paragraph level discourse elements and organization quality. Multi-task learning has been successfully applied in many NLP tasks [Collobert and Weston, 2008]. Our approach is most related to [Søgaard and Goldberg, 2016; Hashimoto *et al.*, 2017; Sanh *et al.*, 2019; Farag and Yannakoudakis, 2019] that arrange multiple tasks in hierarchical structures, while our architecture is specific for modeling organization.

3 Data

3.1 Discourse Elements

The concept of discourse elements is borrowed from [Burstein *et al.*, 2003] and indicates the functions of sentences. In this paper, we refer discourse elements as both sentence functions and paragraph functions.

Sentence Functions. We mainly follow the definition and taxonomy proposed by [Burstein *et al.*, 2003] except that we divide *support* into *evidence* and *elaboration* to give more details. The sentence functions include:

- **Introduction** is to introduce the background or attract readers' attention before making claims.
- **Thesis** expresses the central claim of the writer with respect to the essays topic.
- **Main Idea** asserts foundational ideas or aspects that are related to the thesis.
- **Evidence** indicates examples and other types of evidence that are used to support the main ideas and thesis.
- **Elaboration** further explains the main ideas or evidence, but contains no evidence.
- **Conclusion** summarizes the full essay and echos or extends the central claim.

Paragraph Functions. The function of a paragraph is determined according to the functions of its sentences. We consider the following paragraph functions:

- **IntroductionPara** contains introduction sentences but does not have thesis or main idea sentences.
- **ThesisPara** contains at least a thesis sentence.
- **IdeaPara** contains at least a main idea sentence but does not have a thesis sentence.
- **SupportPara** contains evidence or elaboration sentences but does not contain thesis, main idea or conclusion sentences.
- **ConclusionPara** contains conclusion sentences but does not have thesis sentences.

Basic Statistics	**Number**
#Essays	1,220
Avg. #paragraph per essay	8
Avg. #sentences per essay	28
Avg. #words per sentence	21
Sentence Functions	
Introduction	3,125
Thesis	1,061
Main Idea	4,948
Evidence	6,569
Elaboration	13,351
Conclusion	3,379
Paragraph Functions	
IntroductionPara	893
ThesisPara	864
IdeaPara	3,379
SupportPara	2,788
ConclusionPara	1,796
Organization Grades	
Great	245
Medium	670
Bad	305

Table 1: Basic statistics of the argumentative student essay dataset.

3.2 Organization Grades

We represent organization quality with three grades.

- **Bad** The essay is poorly structured. It is incomplete or misses key discourse elements.
- **Medium** The essay is well structured and complete, but could be further improved.
- **Great** The essay is fairly well structured and the organization is very clear and logical.

3.3 Data Annotation

We built a dataset of argumentative essays written by Chinese high school students. This Chinese corpus could complement the one built by [Persing *et al.*, 2010], which is in English but is not free. Our dataset has 1,220 essays. The topics are diverse and without prompt. We asked two annotators who are high school Chinese teachers to assign function labels to sentences according to an initial manual. After several rounds of practices and revisions, they reached a consensus on the manual. Each essay was labeled by two annotators. We view their annotations as the gold answer and the prediction respectively and the accuracy is 80% and the macro F1 is 77%, indicating the human performance on sentence function identification.

The annotators also assigned organization grades to essays. Again they discussed and revised a manual on the definitions and distinctions of three grades before starting annotation. The inner-rater agreement on the final annotations is 0.73 computed with Kappa.

For both sentence function and organization grade annotation, a third annotator was brought in to discuss with the two annotators on the disagreed annotations to reach a final decision. Table 1 shows the statistics of the dataset.

4 The Proposed Model

Our aim is to evaluate the organization quality of argumentative student essays based on discourse elements. We focus on three tasks: sentence function identification, paragraph function identification and organization evaluation.

Sentence function identification (SFI) is to assign sentence function labels $\mathrm{Y} = y_1, ..., y_n$ to sentences $\mathrm{X} = x_1, ..., x_n$ in an essay, where x_i, $1 \leq i \leq n$, is a sentence of a sequence of words and $y_i \in \mathcal{Y}$, where $\mathcal{Y}$ is the set of pre-defined sentence functions.

Paragraph function identification (PFI) is to assign function labels to m paragraphs $\mathrm{P} = P_1, ..., P_m$.

Organization evaluation (OE) is to give a grade z to an essay, where $z \in \{Bad, Medium, Great\}$.

The three tasks are at different linguistic levels but should affect each other. As shown in Figure 2, we propose a hierarchical neural multi-task learning approach to jointly model these tasks. In addition to the sentence content representation layer, there are three components corresponding to three tasks.

4.1 Sentence Content Representation

The sentence content representations summarize the semantics of sentences and are shared by all the tasks. A sentence of a sequence of words $x = w_1, ..., w_N$ is modeled with a RNN encoder and is converted into a sequence of hidden states $\mathbf{H} = \{\mathbf{h}_1, ..., \mathbf{h}_N\}$. The hidden state at the i-th step is

$$\mathbf{h}_i = f\left(\mathbf{e}(w_i), \mathbf{h}_{i-1}\right), \tag{1}$$

where f is a RNN unit, $\mathbf{e}(w_i) \in \mathbb{R}^{d_w}$ is the embedding of a word, and $\mathbf{h}_{i-1}$ is the hidden state of the previous step. We use Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] as the RNN unit and the sequence is encoded in a bidirectional way that a hidden state $\mathbf{h}_i = [\overrightarrow{\mathbf{h}}_i; \overleftarrow{\mathbf{h}}_i]$ is the concatenation of the corresponding hidden states from both directions. We compute the mean of hidden states $\mathbf{H}$ as the content representation $\mathbf{c}$. The dimension of $\mathbf{c}$ is d.

4.2 Sentence Function Identification Component

In addition to the content representation, we also consider the positional representation of a sentence. This is because the position of a sentence is highly related to some sentence functions.

Positional Representation. We consider three types of sentence positions for positional encoding.

- **Global position**: We view an essay as a sequence of sentences and the global position is the position of the sentence in the sentence sequence.
- **Paragraph position**: Paragraph position refers to the position of the paragraph that contains the sentence in the paragraph sequence.
- **Local position**: Local position is the position of the sentence in its paragraph.

We represent each type of position for the i-th sentence by computing the sinusoidal positional encoding following the transformer model [Vaswani *et al.*, 2017], i.e., $\mathbf{pos}_{global}(i)$,

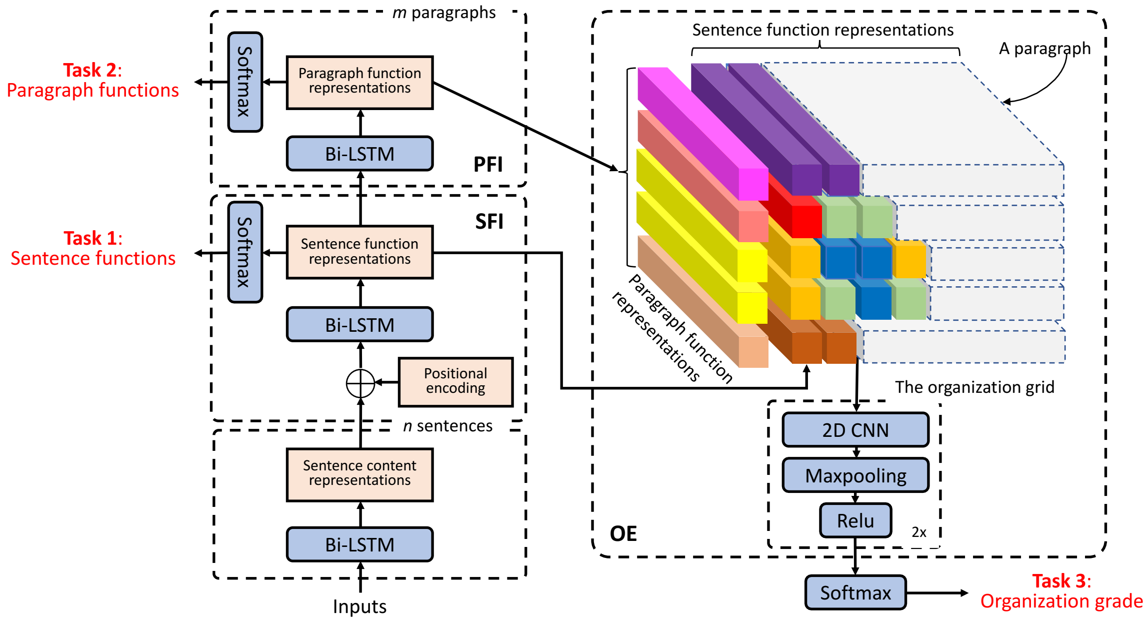

Figure 2: The architecture of the proposed model with sentence function identification (SFI), paragraph function identification (PFI) and organization evaluation (OE) components.

$\mathbf{pos}_{para}(i)$ and $\mathbf{pos}_{local}(i)$. The dimension of the positional encodings is set to d as well. To integrate them together, we use a linear combination to get the final positional encoding:

$$\mathbf{pos}(i) = \sum_{t \in \{global, local, para\}} \beta_t \mathbf{pos}_t(i), \tag{2}$$

where $\{\beta_t\}$ are parameters to be learnt during training.

Function Representation. We use a BiLSTM and a non-linear layer to get the contextual sentence representations:

$$\mathbf{D} = \tanh(\mathrm{BiLSTM}(\mathbf{C} + \mathbf{pos})), \tag{3}$$

where $\mathbf{C} = \{\mathbf{c}_1, ..., \mathbf{c}_n\} \in \mathbb{R}^{d \times n}$ is the content representations of sentences; finally use a linear layer and a $\mathrm{softmax}$ layer to get the probability distributions over sentence function labels for every sentence:

$$\mathbf{Y} = \mathrm{softmax}(\mathrm{linear}(\mathbf{D})). \tag{4}$$

The loss function is the mean of the negative log likelihood over all correct sentence function labels, noted as $\mathcal{L}_{SFI}$.

4.3 Paragraph Function Identification Component

The function of a paragraph is determined by its sentence functions. However, we decide to predict the function of a paragraph rather than deriving it from its sentence functions to enhance interactions between different linguistic levels.

For the l-th paragraph that has sentence indexes from j to k, we feed the corresponding sentence function representations $\mathbf{D}_{[j:k]}$ to a BiLSTM layer. Since the paragraph function is closely related to specific sentence functions, we use the attention mechanism to capture the key sentence functions. The paragraph representation is

$$\mathbf{P}_l = \mathrm{BiLSTM}(\mathbf{D}_{[j:k]}). \tag{5}$$

Again, we use a linear layer and a $\mathrm{softmax}$ layer to get the probability distributions over paragraph function labels for every paragraph. The loss function is the mean of the negative log likelihood over all correct paragraph function labels, noted as $\mathcal{L}_{PFI}$.

4.4 Organization Evaluation Component

We propose to use a grid representation of organization to properly integrate the paragraph function and sentence function representations for organization evaluation.

Organization Grid. Given an essay, we fill a grid $\mathbf{G} \in \mathbb{R}^{m \times (1+n_p)}$ as a part of the input by putting the indexes of sentences and paragraphs at the corresponding coordinates, where m is the maximum number of paragraphs and n_p is the maximum number of sentences in any paragraph. Each row in this grid corresponds to a paragraph. The first column in each row records the paragraph index, followed by a sequence of global indexes of sentences in this paragraph. As a result, the organization grid actually simulates the visual layout of an essay as shown in Figure 2.

In computation, the sentence function representations and the paragraph function representations are read from $\mathbf{D}$ and $\mathbf{P}$ to expand $\mathbf{G}$ to $\mathbf{G}' \in \mathbb{R}^{d \times m \times (1+n_p)}$. The essay organization is finally represented as a tensor.

$\mathbf{G}'$ is fed into two 2dCNN blocks, each of which has multiple filters, a $\mathrm{maxpooling}$ layer and the non-linear function relu. The convolutional operations are done along the grid plane. The feature maps finally are transformed to a vector. A linear layer and a $\mathrm{softmax}$ layer are used for predicting an organization grade as a classification problem. The loss function is the mean of the negative log likelihood over all essay organization grades, noted as $\mathcal{L}_{OE}$.

4.5 The Final Loss Function

The final loss function is

$$\mathcal{L} = \mathcal{L}_{SFI} + \mathcal{L}_{PFI} + \gamma \cdot \mathcal{L}_{OE}. \quad (6)$$

Empirically, we assume SFI and PFI are equally important and use a hyper-parameter γ to control the relative importance of OE. We dynamically update γ in the following way:

$$\gamma = \max(\min(\frac{\mathcal{L}_{OE}}{\mathcal{L}_{SFI}} \cdot \gamma, 1), 0.01). \quad (7)$$

γ is initialized to 0.1 so that the model focuses on optimizing lower level tasks at first and γ becomes larger when $\mathcal{L}_{SFI}$ becomes relatively smaller than $\mathcal{L}_{OE}$.

5 Experiment

5.1 Settings

The sentences are segmented into words. The maximum number of words in a sentence is set to 40. The maximum numbers of sentences, paragraphs and sentences in any paragraph (i.e., n, m and n_p) are set to 50, 20 and 20 empirically. Sentences and paragraphs that are shorter or longer than the limitations are padded or truncated.

We split our dataset into five folds, which have similar distributions over organization grades. Cross-validation was conducted and the average performance would be reported. During training, we randomly selected 10% of the training data as the validation set to find the optimal hyper-parameters.

The dimension of all BiLSTM hidden layers is set to 256. We use the Tencent pre-trained word embeddings for initiation and the dimension is 200 [Song *et al.*, 2018]. The optimizer is stochastic gradient descent (SGD). The two 2dCNN blocks that receives the organization grid have 64 and 32 filters respectively. The kernel size is 5×5.

5.2 Evaluating SFI

This section introduces the evaluation on sentence function identification.

Baselines. We mainly compare with learning based approaches. The heuristic rules proposed by [Persing *et al.*, 2010] are difficult to be adapted or extended for us.

- **Feature-CRF [Song *et al.*, 2015]** This method is based on manually derived features such as *position*, *discourse markers*, *lexical features*, *cohesion* and so on. It uses conditional random fields (CRFs) as the model.
- **HiBiLSTM [Yang *et al.*, 2016]** A hierarchical BiLSTM model is adapted with one attention based BiLSTM layer to encode sentence content and another attention based BiLSTM to encode sentence functions with the content representations as the input.

Our Variants. As introduced in Section 4.2, we use the positional encodings to enhance sentence representations. As a result, our single task model SFI is an enhanced HiBiSTM. In multi-task settings, we could explore different combinations of SFI, PFI and OE components. For example, SFI+OE indicates a combination of SFI and OE components and in this case, the PFI component is removed.

	Systems	Accuracy	Macro F1
Baseline	Feature-CRF	0.629	0.526
	HiBiLSTM	0.626	0.571
Single task	SFI	0.673	0.645
Multi-task	SFI+OE	0.680	0.651
	SFI+PFI	**0.684**	**0.657**
	SFI+PFI+OE	0.680	0.655

Table 2: System comparisons on sentence function identification.

Results. Table 2 shows the system comparison results on SFI. The metrics are accuracy and macro F1. We can see that our single task model SFI outperforms the baselines largely indicating the effectiveness of the positional encoding. SFI+PFI gains the best performance with a significant improvement of 1% on the metrics compared with SFI at $p < 0.05$. SFI+PFI+OE and SFI+OE perform significantly better than SFI at $p < 0.15$. All significance tests are one-tailed paired t-tests.

5.3 Evaluating PFI

As shown in Table 3, the results on PFI show similar trends with the results on SFI. As we introduced in Section 3.1, paragraph functions are determined by sentence functions. As a result, a straightforward method is directly predicting paragraph functions by rules based on the SFI predictions (SFI+Rules) and its performance solely depends on SFI. Jointly training SFI and PFI achieves the best performance. We also try to directly predict paragraph functions with the same architecture as SFI+PFI but removing the output layer of SFI, noted as PFI+OE. As shown in Table 3, its performance is poor, which indicates the importance of supervisions from SFI.

5.4 Evaluating OE

Comparisons. We compare the following systems.

- **[Dong *et al.*, 2017]**: This method is originally proposed for automated essay scoring. It does not consider discourse elements and has only organization grades as supervisions for learning.
- **[Persing *et al.*, 2010]**: This method uses ngrams of paragraph function sequences and sentence function sequences as features and uses SVM to learn a prediction model. We use the multi-task learning model (SFI+PFI) to get sentence and paragraph functions.
- **Our pipelines**: Instead of jointly training three tasks, we train the discourse element identification components and the OE component separately. We first apply SFI+PFI to get sentence and paragraph functions and then train several models for the OE component by receiving different combinations of sentence and paragraph functions as the input.
- **Multi-task learning variants**: We compare three multi-task learning variants. SFI+PFI+OE refers to the complete proposed architecture. SFI+OE removes the PFI

	Systems	Accuracy	Macro F1
Single task	SFI+Rules	0.628	0.585
	PFI	0.605	0.461
Multi-task	SFI+PFI	**0.649**	**0.602**
	PFI+OE	0.582	0.398
	SFI+PFI+OE	**0.649**	0.596

Table 3: Performance on paragraph function identification.

	Systems	Macro F1	MSE	MAE
Baseline	[Dong *et al.*, 2017]	0.474	0.529	0.480
	[Persing *et al.*, 2010]	0.478	**0.386**	0.383
Pipeline	SFI→OE	0.530	0.407	0.394
	PFI→OE	0.468	0.416	0.405
	SFI, PFI→OE	0.524	0.406	0.394
Multitask	SFI+OE	0.565	0.423	0.398
	PFI+OE	0.521	0.480	0.444
	SFI+PFI+OE	**0.591**	0.391	**0.373**

Table 4: Performance on organization evaluation.

Architecture of OE	Macro F1	MSE	MAE
SeqCNN	0.565	0.452	0.409
SeqBiLSTM	0.573	0.421	0.394
GridCNN	**0.591**	**0.391**	**0.373**

Table 5: The effectiveness of the organization grid compared with sequence representations of paragraph and sentence functions.

component and jointly trains SFI and OE so that OE depends on sentence functions only. PFI+OE removes the output layer of SFI but retains the other structures.

Evaluation Metrics. We adopt mean average error (MAE) and mean square error (MSE) as evaluation metrics according to [Persing *et al.*, 2010], where the three grades are mapped to numerical values 0, 1, 2. We also report macro F1 because of the imbalanced distribution over three grades.

Results. Table 4 shows the results. We can see that most of our pipeline and multi-task learning based variants outperform [Dong *et al.*, 2017]. This indicates that considering discourse elements is important. [Persing *et al.*, 2010] achieves very good MSE and MAE scores but low macro F1 score, because it predicts many more *medium* essays but fails to recognize *great* and *poor* essays. In contrast, encoding discourse elements increases the discriminative ability. Considering the pipeline models, SFI is more important than PFI and integrating them gains comparable MSE and MAE, but lower Macro F1 compared with SFI→OE. In the pipeline setting, PFI does not help OE.

In multi-task setting, SFI+PFI+OE achieves the best macro F1 and MAE, and competitive MSE. This indicates that the joint model keeps a better balance on combining the high level abstraction provided by paragraph functions and the details provided by sentence functions. The multi-task learning based models gain better macro F1 scores comparing with their corresponding pipeline models, because they are more effective on identifying minority grades (*great* and *bad*).

The Effectiveness of the Organization Grid. The organization grid organizes sentence functions paragraph by paragraph, noted as **GridCNN**. We compare it with the commonly used sequence representations. First, a CNN or a BiLSTM encoder summarizes the paragraph function sequence and the sentence function sequence into one vector, respectively. Then, the two vectors are concatenated and fed into a linear layer and a $\mathrm{softmax}$ layer for prediction, noted as **SeqCNN** and **SeqBiLSTM**.

Table 5 shows that GridCNN obtains superior performance compared with SeqCNN and SeqLSTM. This indicates that the organization grid is an effective way to represent multiple level discourse elements for organization evaluation.

5.5 Discussions

Based on the above experimental results, we have the following observations. First, organization evaluation can benefit from multi-task learning, indicating that the end-to-end model with multiple level supervisions could avoid suffering from sub-optimal medium representations and error propagation. Second, combining SFI and PFI enhances OE in the multi-task learning setting. Since paragraph function labels are easy to get, the hierarchical architecture can improve model capability at little expense. Third, discourse element identification benefits from multi-task learning. The main contribution comes from the interactions between SFI and PFI. Moreover, the organization grid shows to be effective. The reason may be that the grid representation recovers the visual layout of an essay and captures bigger receptive fields. In addition to sequence patterns in adjacent sentences, relations between sentences over multiple paragraphs could also be captured so that clear and regular structures gain rewards.

6 Conclusion

We presented a hierarchical neural multi-task learning approach for joint discourse element identification and organization evaluation. We have shown that the proposed approach leads to significant improvements on sentence and paragraph level discourse element identification compared with single task models. The improvements mainly come from the mutual enhancement between multiple linguistic levels. The joint model achieves superior performance compared with optimized pipeline models for organization evaluation. It integrates supervisions and representations from lower levels and avoids error propagation. We also demonstrate that the grid representation of the organization of argumentative essays is an effective manner for organization evaluation.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (61876113, 61876112, 2018YFB10 05100), the Beijing Natural Science Foundation (4192017) and Support Project of High-level Teachers in Beijing Municipal Universities in the Period of 13th Five-year Plan (CIT &TCD20170322). Lizhen Liu is the corresponding author.

References

[Attali and Burstein, 2006] Yigal Attali and Jill Burstein. Automated essay scoring with e-rater v. 2. *The Journal of Technology, Learning and Assessment*, 4(3), 2006.

[Burstein *et al.*, 2003] Jill Burstein, Daniel Marcu, and Kevin Knight. Finding the write stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems*, 18(1):32–39, 2003.

[Collobert and Weston, 2008] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167, 2008.

[Dong *et al.*, 2017] Fei Dong, Yue Zhang, and Jie Yang. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162, 2017.

[Farag and Yannakoudakis, 2019] Youmna Farag and Helen Yannakoudakis. Multi-task learning for coherence modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 629–639, 2019.

[Hashimoto *et al.*, 2017] Kazuma Hashimoto, Yoshimasa Tsuruoka, Richard Socher, et al. A joint many-task model: Growing a neural network for multiple nlp tasks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1923–1933, 2017.

[Hearst, 1997] Marti A Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64, 1997.

[Higgins *et al.*, 2004] Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. Evaluating multiple aspects of coherence in student essays. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 185–192, 2004.

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

[Ke *et al.*, 2018] Zixuan Ke, Winston Carlile, Nishant Gurrapadi, and Vincent Ng. Learning to give feedback: Modeling attributes affecting argument persuasiveness in student essays. In *IJCAI*, pages 4130–4136, 2018.

[Mann and Thompson, 1988] William C Mann and Sandra A Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281, 1988.

[Page, 1966] Ellis B Page. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243, 1966.

[Persing *et al.*, 2010] Isaac Persing, Alan Davis, and Vincent Ng. Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239, 2010.

[Sanh *et al.*, 2019] Victor Sanh, Thomas Wolf, and Sebastian Ruder. A hierarchical multi-task approach for learning embeddings from semantic tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6949–6956, 2019.

[Smith, 2003] Carlota S Smith. *Modes of discourse: The local structure of texts*, volume 103. Cambridge University Press, 2003.

[Søgaard and Goldberg, 2016] Anders Søgaard and Yoav Goldberg. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–235, 2016.

[Song *et al.*, 2015] Wei Song, Ruiji Fu, Lizhen Liu, and Ting Liu. Discourse element identification in student essays based on global and local cohesion. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2255–2261, 2015.

[Song *et al.*, 2017] Wei Song, Dong Wang, Ruiji Fu, Lizhen Liu, Ting Liu, and Guoping Hu. Discourse mode identification in essays. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 112–122, 2017.

[Song *et al.*, 2018] Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 175–180, 2018.

[Stab and Gurevych, 2017a] Christian Stab and Iryna Gurevych. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659, 2017.

[Stab and Gurevych, 2017b] Christian Stab and Iryna Gurevych. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990, 2017.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[Wachsmuth *et al.*, 2016] Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 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, 2016.

[Yang *et al.*, 2016] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 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, 2016.