Countering the Influence of Essay Length in Neural Essay Scoring

Sungho Jeon and **Michael Strube**
Heidelberg Institute for Theoretical Studies gGmbH
{sungho.jeon, michael.strube}@h-its.org

Abstract

Previous work has shown that automated essay scoring systems, in particular machine learning-based systems, are not capable of assessing the quality of essays, but are relying on essay length, a factor irrelevant to writing proficiency. In this work, we first show that state-of-the-art systems, recent neural essay scoring systems, might be also influenced by the correlation between essay length and scores in a standard dataset. In our evaluation, a very simple neural model shows the state-of-the-art performance on the standard dataset. To consider essay content without taking essay length into account, we introduce a simple neural model assessing the similarity of content between an input essay and essays assigned different scores. This neural model achieves performance comparable to the state of the art on a standard dataset as well as on a second dataset. Our findings suggest that neural essay scoring systems should consider the characteristics of datasets to focus on text quality.

1 Introduction

Automated essay scoring (AES) is the task of assigning a score for a given essay, aiming to replicate human scoring results. The public release of a standard dataset from a shared task[1] increased the interest in this task significantly. There have been several systems applied to the standard dataset including machine learning-based systems employing diverse features (Chen and He, 2013; Phandi et al., 2015) and neural essay scoring systems (Taghipour and Ng, 2016; Dong et al., 2017).

Previous work, nevertheless, has shown that AES systems are not capable of assessing the quality of essays (Winerip, 2005; Ben-Simon and Bennett, 2007; Wolfe et al., 2016), but indeed work by adopting shallow heuristics for the majority of training examples. Perelman (2014) argues that machine learning-based systems rely on the factor of essay length, and that the high correlation between essay length and scores in the standard dataset leads to top performance. Following this criticism, AES systems must not rely on essay length, a factor irrelevant to writing proficiency (Madnani and Cahill, 2018).

Recent neural essay scoring systems, which do not employ a feature capturing essay length explicitly, achieve state-of-the-art performance. In this work, however, we first show that even neural essay scoring systems might also be influenced by the correlation between essay length and scores in the standard dataset. To investigate this, we here present a simple neural model manipulating essay length. We notice that averaged RNN outputs are the common component in previous neural models, and we modify this component to manipulate essay length. This simple neural model shows performance comparable to state-of-the-art neural models in the standard dataset. However, this artifact does not hold for a second dataset, the Test of English as a Foreign Language dataset (TOEFL, Blanchard et al. (2013)), which has a lower correlation between essay length and scores.

Second, we demonstrate that considering essay content without taking essay length into account can improve the performance of a neural essay scoring system. We incorporate a feature representing Kullback-Leibler divergence into our first simple model, which measures the difference between probability distributions. The intuition is that this feature lets the model consider essay content by assessing the similarity of the word distributions in an input essay and essays assigned different scores. We demonstrate that this neural model achieves performance comparable to the state of the art on both datasets. Our experiments show that neural essay scoring systems might also be influenced by the characteristics of the standard dataset[2].

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

[2] https://github.com/sdeva14/sustai21-counter-neural-essay-length

Proceedings of the 2nd Workshop on Simple and Efficient Natural Language Processing, pages 32–38
November 10, 2021. ©201 Association for Computational Linguistics

2 Essay Length and Scores in Datasets

Datasets: We use two essay datasets, the Automated Student Assessment Prize (ASAP) dataset and the TOEFL dataset, respectively. ASAP was introduced in the shared task on evaluating AES systems, measured against human scores. The essays are written by students in grade levels 7 to 10 of US middle schools. Since the shared task, ASAP has been used as a standard dataset for automatic essay scoring. The dataset consists of eight prompts with different linguistic characteristics such as concreteness vs. open-endedness and different scoring ranges. TOEFL is the standard English test for the entrance to colleges and universities for non-native students. This dataset has not been commonly used for AES, while it has been used as a standard dataset for another shared task, native language identification (Malmasi et al., 2017) (see the supplementary material for details).

Correlation between essay lengths and scores: To uncover relationships between essay length and scores, we check Pearson and Spearman's rank correlation coefficient with p-value < 0.001. In ASAP, all prompts have a high correlation between length and assigned scores: the average of Pearson is 0.702 and the average of Spearman's is 0.707. Only prompt 8 has a rather low correlation between length and the assigned scores. As Perelman (2014) describes, the large range of scores used in prompt 8 (scores range from 1 to 60) causes statistical noise which leads to the low correlation. TOEFL, in general, shows a lower but still high correlation between essay length and scores: the average of Pearson is 0.591 and the average of Spearman's is 0.568.

3 Experimental Setup

Following recent work on ASAP, we evaluate performance at the prompt level in Table 1. We compare with the best performance reported in the literature. Table 2 reports the performance of models on TOEFL achieved by our re-implementations. Our experimental setup is described as follows (see the supplementary material for details).

Implementation Details: While previous neural models deploy pre-trained embeddings for the ASAP dataset (Taghipour and Ng, 2016), our model builds upon Glove, the 100-dimensional pre-trained embedding model trained on Google News (Pennington et al., 2014). We use 100-dimensional Glove for all models on TOEFL. All other settings are identical with previous work.

For our models, we test two variations for an RNN module with a Gated Recurrent Unit (GRU, Cho et al., 2014) and a large-scale natural language pretraining model (XLNet, Yang et al., 2019). XLNet not only outperforms BERT (Devlin et al., 2019) which has led to significant improvements in many NLP tasks, but – unlike BERT – XLNet can also handle any input sequence length, required for our datasets. We encode a whole text at once using the pretrained language model.

Evaluation Details: For ASAP, we perform the experiments in line with prior work (Taghipour and Ng, 2016), including the same cross-validation (CV) partitions, the same evaluation measure, Quadratic Weighted Kappa (QWK), which measures agreement between annotators considering the agreement occurring by chance, and the same loss function, mean squared error. We use the ADAM optimizer with a learning rate of 0.001. For TOEFL, we follow the same setup with the setup of ASAP, but we use a different learning rate of 0.003.

For our implementations, the reported results are obtained by the mean of 10 CV runs with different random seeds. We validate statistical significance by a one-sample t-test with p-value < 0.01.

4 Related Work: Essay Scoring Systems

Taghipour and Ng (2016) introduce a model consisting of a convolutional layer, followed by a recurrent layer, and a mean-over-time layer. The recurrent layer encodes the representation of an essay, the mean-over-time layer then averages the RNN outputs to produce an output vector. Dong et al. (2017) replace the mean-over-time layer by the attention mechanism (Bahdanau et al., 2015). Tay et al. (2018) propose a model that consists of a neural coherence feature and temporal mean pooling, representing the flow of the argument and coherence over time, respectively. They mainly discuss their neural coherence feature, which is composed of a parameterized pair of skipped words. Temporal mean pooling averages RNN outputs. Based on Dong et al. (2017), Nadeem et al. (2019) propose a model considering both the word level and the sentence level with attention between adjacent tokens. Interestingly, we notice that averaged RNN

Model	Prompt								Avg QWK
	1	2	3	4	5	6	7	8	
Phandi et al. (2015)	76.1	60.6	62.1	74.2	78.4	77.5	73.0	61.7	70.5
Taghipour and Ng (2016)	82.1	68.8	69.4	80.5	80.7	81.9	80.8	64.4	76.1
Dong et al. (2017)	82.2	68.2	67.2	81.4	80.3	81.1	80.1	70.5	76.4
Tay et al. (2018)	83.2	68.4	69.5	78.8	81.5	81.0	80.0	69.7	76.4
Cozma et al. (2018)	**84.5**	**72.9**	68.4	**82.9**	**83.3**	**83.0**	80.4	72.9	78.5
Averaging-Length-GRU	80.1	67.5	68.0	79.2	80.5	79.9	79.9	53.6	73.6
Manipulating-Length-GRU	83.7	69.6	68.7	79.2	81.1	80.4	79.8	70.6	76.6
Averaging-Length-XLNet	80.7	69.4	65.4	81.5	79.3	80.7	82.2	73.7	76.6
Manipulating-Length-XLNet	80.8	69.4	66.4	81.6	79.2	80.6	82.2	73.5	76.7
Considering-Content-GRU	84.2	70.8	69.0	79.4	81.5	80.9	80.8	71.2	77.2
Considering-Content-XLNet	82.8	70.6	**69.4**	82.7	80.6	82.0	**83.8**	**76.9**	**78.6**

Table 1: ASAP QWK performance comparison

outputs are commonly used in neural essay scoring systems.

In Table 1, Taghipour and Ng (2016) clearly outperform non-neural essay scoring (Phandi et al., 2015), which uses linguistic features in regression-based machine learning (0.761 > 0.705). Though more recent neural systems introduce more complicated models (Dong et al., 2017; Tay et al., 2018), they only modestly improve the performance (0.764 > 0.761). While the performance of recent neural models is plateauing, a non-neural model which combines a string kernel and word embeddings outperforms previous neural models (Cozma et al., 2018).

5 Influence of Essay Length

A neural model manipulating essay length: To investigate the influence of essay length, we present a simple neural model manipulating essay length, which consists of averaged RNN outputs. Instead of normalizing the sum of RNN outputs by the actual length of an essay, we normalize them by the average of essay lengths in a prompt. The intuition is that normalizing by the average of essay lengths penalizes an essay of shorter length. As an essay has a fewer number of tokens, they have a fewer number of RNN outputs while the same denominator is applied for all essays in the same prompt. This allows a simple RNN model to capture the influence of essay length better.

Results: In ASAP, we first evaluate the performance of the model consisting of averaged RNN outputs with actual essay length, which is commonly used in previous neural models (0.736 < 0.764). We then show that our first neural model manipulating essay length with GRU is comparable with the state-of-the-art neural model (0.766 > 0.764). This demonstrates that the simple neural model manipulating essay length is as powerful as state-of-the-art neural models in the standard dataset. In contrast, a simple neural model relying on the pretrained language model not only shows better performance than previous neural models but also show similar performance against the manipulated model using essay length. It supports the claim of the previous work that a large-scale pretrained language model learns linguistic features from input texts (Warstadt et al., 2020).

In TOEFL, in contrast to ASAP, we observe that the simple neural model which manipulates essay length shows lower performance than the state-of-the-art neural model. Since TOEFL has a lower correlation between essay length and scores than ASAP, we view this as evidence that the performance of previous neural models might be influenced by the correlation of essay length and scores in the target dataset.

6 Countering the Influence

A neural model assessing the similarity of essay content: We propose a neural model which considers essay content by assessing the similarity of word distributions in the input essay and essays grouped into three different levels: low, mid, and high. Grouping scores to three levels enables the model to handle the different score ranges of each dataset. KL divergence for an input essay x is then defined as $KL(p_{lvl}, q) = \sum_x p_{lvl}(x) \log \frac{p_{lvl}(x)}{q(x)}$, where p_{lvl} is a word distribution in the essays grouped to level lvl and q is a word distribution in the input essay. In ASAP where prompts have different score ranges, we define the lower 20% of the scores as the low level, the upper 20% of the scores as the high level, and all others as the mid level. In TOEFL, we define three levels corresponding to the

Model	Prompt								Avg Acc (%)
	1	2	3	4	5	6	7	8	
∗Dong et al. (2017)	69.3	66.5	65.8	66.4	68.9	64.2	67.1	65.7	66.7
∗Nadeem et al. (2019)	58.9	55.8	65.6	61.3	57.8	57.5	52.4	52.8	57.8
Averaging-Length-GRU	65.7	65.0	63.0	62.9	66.5	64.4	63.2	62.6	64.2
Manipulating-Length-GRU	65.7	65.1	62.6	63.3	66.4	63.0	63.2	62.6	64.0
Averaging-Length-XLNet	73.3	73.6	69.1	70.6	74.1	72.3	71.5	70.5	71.9
Manipulating-Length-XLNet	73.3	73.6	69.1	70.6	73.8	71.9	71.5	70.4	71.8
Considering-Content-GRU	69.4	67.5	66.3	65.5	69.4	65.8	67.4	64.0	66.9
Considering-Content-XLNet	**74.4**	**74.2**	**70.4**	**71.9**	**74.6**	**72.5**	**72.6**	**71.8**	**72.8**

Table 2: TOEFL Accuracy performance comparison (∗: our re-implementation)

scores in the dataset. Given an essay, we therefore compute three scalar values of KL divergence, and normalize them by the average of KL divergence in the same level. Finally, we concatenate the three scalar values of KL divergence to the vector representing averaged RNN outputs, which is our first simple model.

Results: We evaluate the performance of the model considering essay content using KL divergence. In ASAP, this model with GRU leads to 0.8% QWK improvement compared to the previous neural models ($0.772 > 0.764$). Finally, this model with XLNet shows performance comparable to the non-neural state of the art ($0.786 > 0.785$).

We observe a performance gain with the model considering essay content mainly in prompt 7 and 8, while the non-neural state of the art outperforms this model in prompt 4, 5, and 6. We suspect that the different improvements of performance are caused by different linguistic properties in the essays responding to those prompts. The string kernel method is based on character similarity (Cozma et al., 2018). It has an advantage in a task which is similar to extractive summarization such as prompt in 4, 5, and 6. These prompts ask the students to write an essay within a given context. In this case, we suspect that including specific information regarding the given context has more influence to human annotators. In contrast, prompts 7 and 8 do not have a given context. They are open-ended. Our experiments show that considering essay content leads to significant improvement in this case.

In TOEFL, the model considering essay content with GRU shows performance comparable to the state-of-the-art neural models ($0.669 > 0.667$). Eventually, this model with XLNet outperforms all other models ($0.728 > 0.669$), and it also leads to a 1.0% improvement compared to the model manipulating essay length. This also supports our finding that considering essay content leads to a performance improvement for neural essay scoring systems. Unlike ASAP, the model considering essay content models on the TOEFL show consistent performance gain regardless of prompts. We believe this is caused by an overall higher quality of TOEFL dataset. The prompts do not vary so much, the student population is more controlled, and the essays have a similar length.

We also compare with the state of the art on TOEFL, Nadeem et al. (2019). We notice that the performance reported in Nadeem et al. (2019) cannot be compared with previous work due to a different experimental setup. They filter out content whose sentence length is longer than 40 words or whose document length is longer than 25 sentences, which results in filtering the more than 7.5% of sentences; they also evaluate performance without CV[3]. To ensure a fair comparison, we only modified the experimental setup in their implementation. The model proposed in Nadeem et al. (2019) shows substantially lower performance than all models in the same experimental setup with previous work.

7 Conclusions

While the recent neural essay scoring systems emphasize novel aspects of the neural architecture, they have neither shown significant improvement nor helped interpreting the scores assigned by humans. We show that recent neural models might also be influenced by characteristics of the standard dataset, a high correlation between essay length and scores. To investigate this, we evaluate models on two datasets, not only the standard dataset, ASAP, but also TOEFL, which has a lower correlation between essay length and scores. Our findings suggest that neural essay scoring systems should focus on text quality, and at the same time, should consider characteristics of the target dataset.

[3] We confirmed this by examining their implementation and emailing the first author.

Acknowledgments

The authors would like to thank the anonymous reviewers for their comments. This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany. The first author has been supported by a Heidelberg Institute for Theoretical Studies Ph.D. scholarship.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *In Proceedings of the ICLR Conference*.

Anat Ben-Simon and Randy Elliot Bennett. 2007. Toward more substantively meaningful automated essay scoring. *The Journal of Technology, Learning and Assessment*, 6(1).

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.

Hongbo Chen and Ben He. 2013. Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1741–1752, Seattle, Washington, USA. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Mădălina Cozma, Andrei Butnaru, and Radu Tudor Ionescu. 2018. Automated essay scoring with string kernels and word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 503–509, Melbourne, Australia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fei Dong, Yue Zhang, and Jie Yang. 2017. 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, Vancouver, Canada. Association for Computational Linguistics.

Nitin Madnani and Aoife Cahill. 2018. Automated scoring: Beyond natural language processing. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1099–1109, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel Tetreault, Robert Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. 2017. A report on the 2017 native language identification shared task. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 62–75, Copenhagen, Denmark. Association for Computational Linguistics.

Farah Nadeem, Huy Nguyen, Yang Liu, and Mari Ostendorf. 2019. Automated essay scoring with discourse-aware neural models. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 484–493, Florence, Italy. Association for Computational Linguistics.

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, Doha, Qatar. Association for Computational Linguistics.

Les Perelman. 2014. When “the state of the art” is counting words. *Assessing Writing*, 21:104–111.

Peter Phandi, Kian Ming A. Chai, and Hwee Tou Ng. 2015. Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 431–439, Lisbon, Portugal. Association for Computational Linguistics.

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, Austin, Texas. Association for Computational Linguistics.

Yi Tay, Minh C Phan, Luu Anh Tuan, and Siu Cheung Hui. 2018. Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020. Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). In *Proceedings of the 2020 Conference on Empirical Methods* *in Natural Language Processing (EMNLP)*, pages 217–235, Online. Association for Computational Linguistics.

Michael Winerip. 2005. SAT essay test rewards length and ignores errors. *New York Times*, 9.

Edward W. Wolfe, Tian Song, and Hong Jiao. 2016. Features of difficult-to-score essays. *Assessing Writing*, 27:1–10.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, pages 5754–5764.

8 Appendix A. Dataset Details

Table 3 describes statistics on two datasets, GCDC[4] and TOEFL[5]. We use NLTK library to tokenize for models based on GRU, and use XLNet tokenizer for the models based on XLNet. Table 4 describes the topic of each prompt in TOEFL. They are all open-ended tasks, that do not have given context but require students to submit their opinion.

Dataset	#Texts	Avg len (Std)	Max len	Scores
A-P1	1,785	463 (155)	1092	2-12
A-P2	1,800	467 (194)	1337	1-6
A-P3	1,726	135 (67)	452	0-3
A-P4	1,772	114 (64)	451	0-3
A-P5	1,805	153 (73)	520	0-4
A-P6	1,800	187 (69)	545	0-4
A-P7	1,569	223 (119)	878	0-30
A-P8	723	770 (269)	1396	0-60
T-P1	1,656	401 (97)	902	1-3
T-P2	1,562	423 (97)	902	1-3
T-P3	1,396	407 (102)	837	1-3
T-P4	1,509	405 (99)	852	1-3
T-P5	1,648	424 (101)	993	1-3
T-P6	960	425 (101)	925	1-3
T-P7	1,686	396 (87)	755	1-3
T-P8	1,683	407 (92)	795	1-3

Table 3: Dataset statistics on tokenization: each ASAP prompt (A-P) and each TOEFL prompt (T-P).

9 Appendix B. Experimental Setup Details

For ASAP, we perform the experiments in line with prior work (Taghipour and Ng, 2016), including the same cross-validation (CV) partitions, the same evaluation measure, Quadratic Weighted Kappa (QWK), which measures agreement between annotators considering the agreement occurring by chance, and the same loss function, mean squared error. We use the ADAM optimizer with a learning rate of 0.001. We evaluate performance for 50 epochs on the validation set. We evaluate performance on the validation set for every epoch and apply the best model to the test set. We use a mini-batch size of 32 with random-shuffle. For TOEFL, we evaluate performance in accuracy for the three-class classification problem with 5-fold CV. We deploy cross-entropy loss for training. Like ASAP, we evaluate performance on the validation set for every epoch and apply the best model to the test set. We use a mini-batch size of 32 with random-shuffle. We use the ADAM optimizer with a learning rate of 0.003. We use 23GB GPU memory a NVidia P40.

[4] https://github.com/aylai/GCDC-corpus

[5] https://catalog.ldc.upenn.edu/LDC2014T06

A-Prompt 1	The writers had to write a letter to their local newspaper in which they stated their opinion on the effects computers have on people.
A-Prompt 2	The writers had to write a persuasive essay reflecting their views on censorship in libraries.
A-Prompt 3	The writers had to read an extract from Rough Road Ahead: Do Not Exceed Posted Speed Limit by Joe Kurmaskie. They then had to explain how the features of the setting affected the cyclist.
A-Prompt 4	The writers had to read an extract from Winter Hibiscus by Minfong Ho. They then had to explain why the author concludes the story in the way that she did.
A-Prompt 5	The writers had to read an extract from Narciso Rodriguez by Narciso Rodriguez. They then had to describe the mood created by the author with supporting evidence from the extract.
A-Prompt 6	The writers had to read an extract from The Mooring Mast by Marcia Amidon Lusted. They then had to answer a question about the difficulties faced by the builders of the Empire State Building in allowing dirigibles to dock there.
A-Prompt 7	Write a story about a time when you, or someone you know, was patient
A-Prompt 8	Write a story in which laughter plays a part.
T-Prompt 1	Agree or Disagree: It is better to have broad knowledge of many academic subjects than to specialize in one specific subject.
T-Prompt 2	Agree or Disagree: Young people enjoy life more than older people do.
T-Prompt 3	Agree or Disagree: Young people nowadays do not give enough time to helping their communities.
T-Prompt 4	Agree or Disagree: Most advertisements make products seem much better than they really are.
T-Prompt 5	Agree or Disagree: In twenty years, there will be fewer cars in use than there are today.
T-Prompt 6	Agree or Disagree: The best way to travel is in a group led by a tour guide.
T-Prompt 7	Agree or Disagree: It is more important for students to understand ideas and concepts than it is for them to learn facts.
T-Prompt 8	Agree or Disagree: Successful people try new things and take risks rather than only doing what they already know how to do well.

Table 4: Topic description: ASAP (A) and TOEFL (T).