References

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

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, August. Association for Computational Linguistics.

Camille Guinaudeau and Michael Strube. 2013. Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 93–103, Sofia, Bulgaria, August. Association for Computational Linguistics.

Alice Lai and Joel Tetreault. 2018. Discourse coherence in the wild: A dataset, evaluation and methods. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 214–223, Melbourne, Australia, July. Association for Computational Linguistics.

Jiwei Li and Dan Jurafsky. 2017. Neural net models of open-domain discourse coherence. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 198–209, Copenhagen, Denmark, September. Association for Computational Linguistics.

Mohsen Mesgar and Michael Strube. 2018. A neural local coherence model for text quality assessment. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4328–4339, Brussels, Belgium, October-November. 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, August. Association for Computational Linguistics.

1 Appendix A. Dataset Details

Table 1 describes statistics on two datasets, GCDC[1] and TOEFL[2]. We split a text at the sentence level by Stanford Stanza library, and tokenize them by the XLNet tokenizer. Table 2 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
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
G-Y	1,200	173 (48)	378	1-3
G-C	1,200	200 (65)	385	1-3
G-E	1,200	203 (67)	388	1-3
G-P	1,200	198 (58)	374	1-3

Table 1: Dataset statistics on tokenization: each TOEFL prompt (T-P) and four domains in GCDC, Yahoo (G-Y), Clinton (G-C), Enron (G-E), and Yelp (G-P)

2 Appendix B. Experiments Detail

[1]https://github.com/aylai/GCDC-corpus
[2]https://catalog.ldc.upenn.edu/LDC2014T06

Prompt 1	Agree or Disagree: It is better to have broad knowledge of many academic subjects than to specialize in one specific subject.
Prompt 2	Agree or Disagree: Young people enjoy life more than older people do.
Prompt 3	Agree or Disagree: Young people nowadays do not give enough time to helping their communities.
Prompt 4	Agree or Disagree: Most advertisements make products seem much better than they really are.
Prompt 5	Agree or Disagree: In twenty years, there will be fewer cars in use than there are today.
Prompt 6	Agree or Disagree: The best way to travel is in a group led by a tour guide.
Prompt 7	Agree or Disagree: It is more important for students to understand ideas and concepts than it is for them to learn facts.
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 2: Topic description: TOEFL

Model	Prompt								Avg Acc
	1	2	3	4	5	6	7	8	
Dong et al. (2017)	0.693	0.665	0.658	0.664	0.689	0.642	0.671	0.657	0.667
Mesgar and Strube (2018)	0.549	0.564	0.524	0.561	0.553	0.555	0.560	0.573	0.555
Nadeem et al. (2019)	0.589	0.558	0.656	0.613	0.578	0.575	0.524	0.528	0.578
Avg-GRU	0.657	0.637	0.647	0.656	0.671	0.649	0.651	0.630	0.650
Our Model-GRU	0.659	0.639	0.641	0.655	0.684	0.652	0.647	0.638	0.652
Avg-XLNet	0.742	0.735	0.729	0.727	0.760	0.749	0.729	0.719	0.736
Our Model-XLNet	**0.748**	**0.741**	**0.728**	**0.734**	**0.761**	**0.765**	**0.735**	**0.714**	**0.741**

Table 3: TOEFL Accuracy performance comparison

Model	Yahoo	Clinton	Enron	Yelp	Avg Acc
Barzilay and Lapata (2008)	38.0	43.0	46.0	45.5	43.1
Guinaudeau and Strube (2013)	40.0	56.0	43.5	53.0	48.1
Li and Jurafsky (2017)	53.5	61.0	54.4	49.1	51.7
Mesgar and Strube (2018)	47.3	57.7	50.6	54.6	52.6
Lai and Tetreault (2018)	54.9	60.2	53.2	54.4	55.7
Avg-GRU	54.0	62.4	55.9	55.5	56.9
Our Model-GRU	54.2	63.2	56.3	55.9	57.4
Avg-XLNet-base	61.2	66.1	56.5	58.9	60.7
Our Model-XLNet	**61.4**	**66.1**	**56.4**	**59.2**	**60.8**

Table 4: GCDC Accuracy performance comparison