BabyLM Challenge: Curriculum learning based on sentence complexity approximating language acquisition

Miyu Oba[1] **Akari Haga**[1] **Akiyo Fukatsu**[2] **Yohei Oseki**[2]
[1]Nara Institute of Science and Technology
[2]The University of Tokyo
{oba.miyu.ol2, haga.akari.ha0}@is.naist.jp
{akiyofukatsu, oseki}@g.ecc.u-tokyo.ac.jp

Abstract

This paper describes our proposed models in the BabyLM Challenge (Warstadt et al., 2023). The goal of this shared task is to pretrain models efficiently using a developmentally plausible corpus. To simulate the increasing complexity of Child-Directed Speech (CDS) sentences, we employed curriculum learning and trained models with data reordered based on three metrics for sentence complexity. Among all the models, the best performing one was trained with data ordered by the max-dependency, although the models trained with curriculum learning did not outperform the baseline model without curriculum learning.

1 Introduction

Successful recent large language models (LLMs) are trained on extensive datasets, leading to a gap between the training data of models and the inputs that children receive during language acquisition. English-speaking children hear less than 100M words until the age of 12, while Chinchilla, one of the recent LLMs, uses 1.4 trillion words for training (Wertz et al., 2022). Training models with human-like input data can improve LLM data efficiency and shed light on efficient language acquisition in children with limited data. Thus, the BabyLM Challenge (Warstadt et al., 2023) aims to pretrain models on a developmentally plausible corpus, including Age-Ordered CDS (Huebner and Willits, 2021). We used a dataset of $\sim$10M words, approximating the input that children receive until 2–3 years [1].

In model training, reordering data in a meaningful way (e.g., from easy to difficult samples), known as curriculum learning (Bengio et al., 2009), is suggested to enhance performance. In human language acquisition, mothers adjust their speech when addressing their children, using shorter and simpler sentences (Snow, 1972; Newport et al., 1977; Fernald et al., 1989). Notably, Snow (1972) and Fernald et al. (1989) report that the mean length of utterance and the use of nominal compounds increase as children age, suggesting that language-acquiring children receive easy inputs initially and gradually encounter more complexity as they grow. Thus, reordering data by sentence difficulty may improve model performance.

In this paper, we train models on data reordered by sentence difficulty and evaluate them on three designated datasets. The difficulty metrics include the number of subword tokens, that of constituents and max-dependency. The max-dependency yielded the highest scores, but curriculum learning did not outperform the baseline model.

2 Corpora and preprocessing

We used the BabyLM strict-small train/dev dataset (Warstadt et al., 2023). First, we split the corpora into sentences using the sentencizer from spaCy[2]. Next, we deleted sentences that were non-English, titles, and longer than 300 characters. For identifying non-English sentences, we used FastText (Joulin et al., 2017). Some corpora in the datasets contain much upper-case-only or lower-case-only data. Therefore we trained Moses truecaser (Koehn et al., 2007) using other training corpora, then true-cased all data. After true-casing, we tokenized all data. We trained the tokenizer from scratch using RobertaTokenizer (Liu et al., 2019) with the preprocessed training dataset.

3 Models

3.1 Baseline model

Our models are based on the RoBERTa-base (Liu et al., 2019). We trained them on randomly shuf-

[1]According to Gilkerson et al. (2017), children are exposed to adult 12,300 words within a 12-hour day.

[2]https://spacy.io

Proceedings of the 27th Conference on Computational Natural Language Learning: Volume 2: The BabyLM Challenge, pages 290–297
December 6-7, 2023

fled data from scratch. Their hyperparameters are shown in Appendix A.2.

3.2 Curriculum learning model

We employed curriculum learning in our baseline models. Training data were sorted by a particular difficulty metric. We focused on sentence complexity and used three metrics, the number of subword tokens (Ntoken), that of constituency (Nconst.), and maximum depth of dependency tree (Max-dep.). We split the data into several blocks and trained models on them in order with particular steps. Note that we adjusted the number of steps in each block to be proportional to the number of subwords in each block.

4 Experiments

To find optimal settings for curriculum learning, we begin with investigating which difficulty metrics are better and how many blocks of data should be split into for this task. To explore the effect of curriculum learning, we then compare the baseline model, which is trained on randomly shuffled data, with the curriculum learning models. We use parsers from spaCy to calculate the number of constituents and max-dependency.

4.1 Evaluation

We evaluated our models with the shared evaluation datasets (Gao et al., 2021). These consist of BLiMP (Warstadt et al., 2020a), (Super)GLUE (Wang et al., 2018) and MSGS (Warstadt et al., 2020b). BLiMP is used for zero-shot evaluation, and it includes supplement tasks that are specifically made for BabyLM. We report its accuracy. GLUE and MSGS are used for fine-tuning evaluation. We report F1 score for GLUE and Matthews Correlation Coefficient (MCC) for MSGS.

4.2 Results

Difficulty metrics We compare the models trained on the sorted data with the three difficulty metrics (See section 3.2). The bottom of Table 1 shows the performance of curriculum learning models in the different difficulty metrics. The results suggest that the difficulty metrics affect the performance of the models. Notably, the model trained on the data sorted by Max-dep. achieved slightly higher performance than the other metrics.

Model	Curr.	BLiMP	GLUE	MSGS	Avg.
Baseline		69.23	65.74	−0.57	44.80
+cleaning		**70.46**	**66.40**	**6.86**	**47.91**
Ntoken	✓	**68.37**	64.96	−5.56	42.59
Nconst.	✓	65.90	64.71	−2.73	42.63
Max-dep.	✓	68.27	**65.90**	**3.26**	**45.81**

Table 1: Performance of models. The models at the top are baseline models with and without data preprocessing. Those at the bottom are curriculum learning models in different difficulty metrics. ✓in Curr. denotes whether curriculum learning is applied to the models.

Model	n	BLiMP	GLUE	MSGS	Avg.
Max-dep.	3	**68.70**	65.06	0.37	44.71
	4	68.27	**65.90**	3.26	45.81
	6	67.85	64.97	**9.56**	**47.46**
	8	67.93	65.05	0.33	44.44

Table 2: Performance of models with different split blocks. n indicates the number of blocks.

Number of blocks We compare the models trained on the data split into {3, 4, 6, 8} blocks. As difficulty metrics, we use Max-dep., which achieves the highest score among the three models at the bottom of Table 1. Table 2 indicates the performance of models with different split blocks. This result shows that there is no significant difference between the models with different split blocks, suggesting that scores will not be improved by the simple increase or decrease in the number of split blocks.

Baseline model vs. Curriculum learning model Finally, we compare the curriculum learning model [3], in which difficulty metrics are Max-dep. and the number of blocks is 4, with the baseline model. The top of Table 1 shows that the baseline model obtains higher scores than the curriculum learning model. This result implies that at least the curriculum learning settings attempted in this work are inadequate in facilitating higher model performance. Investigating other effective training settings would be interesting for future work; e.g., warmup, optimizers.

5 Conclusion

In summary, our participation in the BabyLM Challenge centered on curriculum learning based on the three metrics of sentence complexity. While

[3]The model is available at https://huggingface.co/akari000/roberta-dependency-max-4split

the max-dependency demonstrated slightly higher performance scores than the other metrics, it did not outperform the baseline model without curriculum learning on the BLiMP dataset. These findings suggest the complexity of language acquisition and the need to improve the experimental setting in future research to enhance the models' performance. To enhance the validity of our research as a future work, we need to use multiple random seeds to train the model to verify how much those affect the results.

Acknowledgements

We would like to express our gratitude for the anonymous reviewers who provided many insightful comments that have improved our paper. This work was supported by JSPS KAKENHI Grant Number JP21H05054 and JST PRESTO Grant Number JPMJPR21C2.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

Anne Fernald, Traute Taeschner, Judy Dunn, Mechthild Papousek, Bénédicte de Boysson-Bardies, and Ikuko Fukui. 1989. A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Jill Gilkerson, Jeffrey A. Richards, Steven F. Warren, Judith K. Montgomery, Charles R. Greenwood, D. Kimbrough Oller, John H. L. Hansen, and Terrance D. Paul. 2017. Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, 26(2):248–265.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of EMNLP*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Philip A. Huebner and Jon A. Willits. 2021. Chapter eight - using lexical context to discover the noun category: Younger children have it easier. In Kara D. Federmeier and Lili Sahakyan, editors, *The Context of Cognition: Emerging Perspectives*, volume 75 of *Psychology of Learning and Motivation*, pages 279–331. Academic Press.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of EACL: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of ACL (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Elissa. Newport, Henry Gleitman, and Lila Gleitman. 1977. Mother, id rather do it myself: Some effects and non-effects of maternal speech style. In Catherine E. Snow and Charles A. Ferguson, editors, *Talking to Children*, pages 109–149. Cambridge University Press.

Joakim Nivre and Jens Nilsson. 2005. Pseudo-projective dependency parsing. In *Proceedings of ACL*, pages 99–106, Ann Arbor, Michigan. Association for Computational Linguistics.

Catherine E. Snow. 1972. Mothers' speech to children learning language. *Child Development*, 43(2):549–565.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Warstadt, Leshem Choshen, Ryan Cotterell, Tal Linzen, Aaron Mueller, Ethan Wilcox, Williams Adina, and Chengxu Zhuang. 2023. Findings of the BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020a. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

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

Lukas Wertz, Katsiaryna Mirylenka, Jonas Kuhn, and Jasmina Bogojeska. 2022. Investigating active learning sampling strategies for extreme multi label text classification. In *Proceedings of LREC*, pages 4597–4605, Marseille, France. European Language Resources Association.

A Appendix

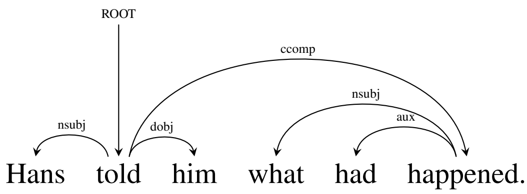
A.1 Difficulty Metrics

Number of constituents The number of constituents was counted using the Berkley Neural Parser (Kitaev and Klein, 2018) in spaCy. This parser uses a self-attentive encoder in place of LSTM along with a chart decoder. This parser outputs POS tags and surface strings in brackets as in (1), and we count the number of phrasal nodes (e.g., NP) in the outputs. In this case, the number of constituents is counted as 4.

(1) (S (NP (DT That)) (VP (MD might) (VP (VB be) (ADJP (JJR better)))) (. .))

Max-dependency We count max-dependency using the dependency parser in spaCy, which is a transition-based system by Honnibal and Johnson (2015) along with Nivre and Nilsson (2005)'s pseudo-projective dependency transformation. We count the number of dependent nodes from the root and choose the maximum depth as the value of max-dependency. For example, the dependency tree in (2) is an example of parsing by the dependency parser. In this case, the longest dependency is either 'told → happened → had' or 'told → happened → what'. Given that the root is counted as 0, the max-dependency of this sentence is 2.

(2)



A.2 Hyperparameters

We arranged the number of instances that we input into our models for all steps to 28,800k instances. Other hyperparameters are shown in Table 3.

A.3 Detailed results

We show the details of the results for each task. Table 4 – 6 shows the accuracies for all measures in BLiMP and GLUE. Table 7 shows the F1 scors for all measures in GLUE, where we use macro-F1, and Table 8 shows the MCC scores for all measures in MSGS.

Model	architecture	roberta-base
	vocab size	50,265
	hidden size	768
	heads	12
	layers	12
	dropout	0.1
	layer norm eps	1e-12
Optimizer	algorithm	AdamW
	learning rates	3e-4
	betas	(0.9, 0.999)
	weight decay	0.1
	clip norm	0.0
Scheduler	type	cosine
	warmup updates	5000
Training	gradient accumulation	4
	line by line	true
	NGPU	4

Table 3: Hyperparameters of the models

Model	Curr.	n	Anaphor Agr.	Agr. Structure	Binding	Control/Raising	D-N Agr.	Ellipsis
Baseline model		-	86.09	73.68	67.84	68.03	95.57	73.44
+cleaning		-	91.82	74.32	74.16	73.75	96.29	77.19
Ntoken	✓	4	88.45	75.24	73.67	73.75	95.47	74.19
Nconst.	✓	4	83.44	72.50	73.75	71.74	91.45	75.17
Max-dep.	✓	3	90.85	73.82	73.76	72.45	95.62	79.68
	✓	4	91.21	74.98	73.49	71.06	95.48	78.58
	✓	6	87.68	71.59	73.79	68.43	93.86	76.21
	✓	8	91.26	72.25	73.43	67.21	94.55	74.54

Model	Curr.	n	Filler Gap	Irregular Forms	Island Effects	NPI Licensing	Quantifiers	S-V Agr.
Baseline model		-	75.26	90.69	37.56	52.73	74.86	78.14
+cleaning		-	76.39	90.99	44.96	56.71	73.98	82.48
Ntoken	✓	4	74.93	89.57	38.08	55.10	72.41	81.43
Nconst.	✓	4	77.65	74.66	39.57	61.75	65.10	76.50
Max-dep.	✓	3	71.49	87.48	35.24	57.91	72.05	81.70
	✓	4	71.88	88.80	33.15	53.72	71.95	83.04
	✓	6	71.94	89.87	26.76	60.04	69.91	81.43
	✓	8	72.08	91.40	28.70	58.56	75.94	78.34

Table 4: Accuracies for all measures in BLiMP

Model	Curr.	n	Hypernym	QA Congruence (easy)	QA Congruence (tricky)	Subj.-Aux. Inversion	Turn Taking
Baseline model		-	49.53	60.94	43.03	84.24	65.36
+cleaning		-	49.19	67.19	39.39	68.72	60.36
Ntoken	✓	4	48.72	59.38	40.00	64.85	57.14
Nconst.	✓	4	48.02	54.69	29.70	67.14	57.50
Max-dep.	✓	3	48.84	68.75	36.97	63.09	58.21
	✓	4	46.98	65.63	39.39	63.77	57.50
	✓	6	47.91	67.19	43.03	63.53	60.36
	✓	8	51.40	60.94	37.58	63.38	63.21

Table 5: Accuracies for all measures in BLiMP supplement task

Model	Curr.	n	CoLA	SST-2	MRPC	QQP	MNLI	MNLI-mm
Baseline model		-	72.91	87.01	64.97	80.61	70.04	71.13
+cleaning		-	76.84	88.39	69.49	82.32	72.19	74.06
Ntoken	✓	4	76.15	87.60	64.41	82.31	72.14	71.79
Nconst.	✓	4	73.01	87.01	66.67	82.74	70.41	72.14
Max-dep.	✓	3	75.17	87.40	70.62	83.46	72.90	73.01
	✓	4	75.17	87.20	67.23	82.75	72.37	73.50
	✓	6	75.47	87.60	70.06	82.13	72.04	73.98
	✓	8	75.47	88.39	66.67	83.14	71.96	73.22

Model	Curr.	n	QNLI	RTE	BoolQ	MultiRC	WSC
Baseline model		-	69.25	51.52	65.15	60.35	61.45
+cleaning		-	71.26	52.53	66.67	58.71	63.86
Ntoken	✓	4	66.01	52.53	65.98	59.26	61.45
Nconst.	✓	4	64.92	56.57	66.11	59.15	61.45
Max-dep.	✓	3	71.00	48.48	66.11	60.46	61.45
	✓	4	70.25	52.53	65.98	61.34	61.45
	✓	6	70.73	46.46	64.04	59.04	61.45
	✓	8	70.21	57.58	66.39	59.26	61.45

Table 6: Accuracies for all measures in GLUE task

Model	Curr.	n	CoLA	SST-2	MRPC	QQP	MNLI	MNLI-mm
Baseline model		-	82.92	87.36	74.80	76.27	-	-
+cleaning		-	84.58	88.45	80.58	79.78	-	-
Ntoken	✓	4	83.77	87.52	76.92	79.10	-	-
Nconst.	✓	4	82.22	87.36	77.90	79.36	-	-
Max-dep.	✓	3	83.96	87.64	80.88	80.38	-	-
	✓	4	83.77	87.67	78.68	79.96	-	-
	✓	6	83.66	87.67	80.87	79.12	-	-
	✓	8	83.85	88.54	78.07	79.93	-	-

Model	Curr.	n	QNLI	RTE	BoolQ	MultiRC	WSC
Baseline model		-	72.89	45.45	74.65	57.31	20.00
+cleaning		-	74.47	47.19	76.11	54.63	11.76
Ntoken	✓	4	71.36	52.53	73.72	59.74	00.00
Nconst.	✓	4	71.44	59.05	75.57	49.53	00.00
Max-dep.	✓	3	74.82	45.16	75.52	57.18	00.00
	✓	4	74.46	53.47	74.11	60.99	00.00
	✓	6	72.16	51.38	74.61	55.26	00.00
	✓	8	72.51	55.32	75.92	51.31	00.00

Table 7: F1 scores for all measures in GLUE task

Model	Curr.	n	CR (Control)	LC (Control)	MV (Control)	RP (Control)	SC (Control)	
Baseline model		-	64.29	99.98	92.47	75.34	73.65	
+cleaning		-	76.34	100.00	99.64	99.91	27.19	
Ntoken	✓	4	66.43	100.00	96.66	90.15	53.17	
Nconst.	✓	4	64.76	100.00	97.11	96.48	49.27	
Max-dep.	✓	3	81.61	100.00	99.59	99.98	24.81	
	✓	4	77.71	100.00	99.23	100.00	52.80	
	✓	6	67.47	100.00	99.37	92.35	74.47	
	✓	8	55.99	100.00	98.98	99.82	38.54	
Model	**Curr.**	**n**	**CR_LC**	**CR_TP**	**MV_LC**	**MV_RTP**	**SC_LC**	**SC_RP**
Baseline model		-	-70.37	-69.93	-100.00	-81.71	-57.74	-32.27
+cleaning		-	33.37	-65.21	-99.54	-79.93	-59.83	-56.48
Ntoken	✓	4	-92.54	-44.48	-100.00	-89.32	-78.91	-62.35
Nconst.	✓	4	-47.57	-98.28	-98.55	-85.35	-52.81	-55.07
Max-dep.	✓	3	-39.21	-73.38	-100.00	-83.32	-48.79	-57.24
	✓	4	-32.06	-62.60	-100.00	-77.70	-59.96	-61.52
	✓	6	20.13	-65.46	-100.00	-86.16	-32.50	-64.47
	✓	8	-17.58	-63.82	-100.00	-99.03	-47.53	-61.69

Table 8: MCC scores for all measures in MSGS

Models	Curr.	n	Perplexity
Baseline model		–	14.58
+cleaning		–	19.80
Ntoken	✓	4	25.74
Nconst.	✓	4	32.20
Max-dep.	✓	3	24.42
	✓	4	27.35
	✓	6	38.61
	✓	8	40.16

Table 9: Perplexity for all measures