DACL: Disfluency Augmented Curriculum Learning for Fluent Text Generation

Rohan Chaudhury, Maria Teleki, Xiangjue Dong, James Caverlee

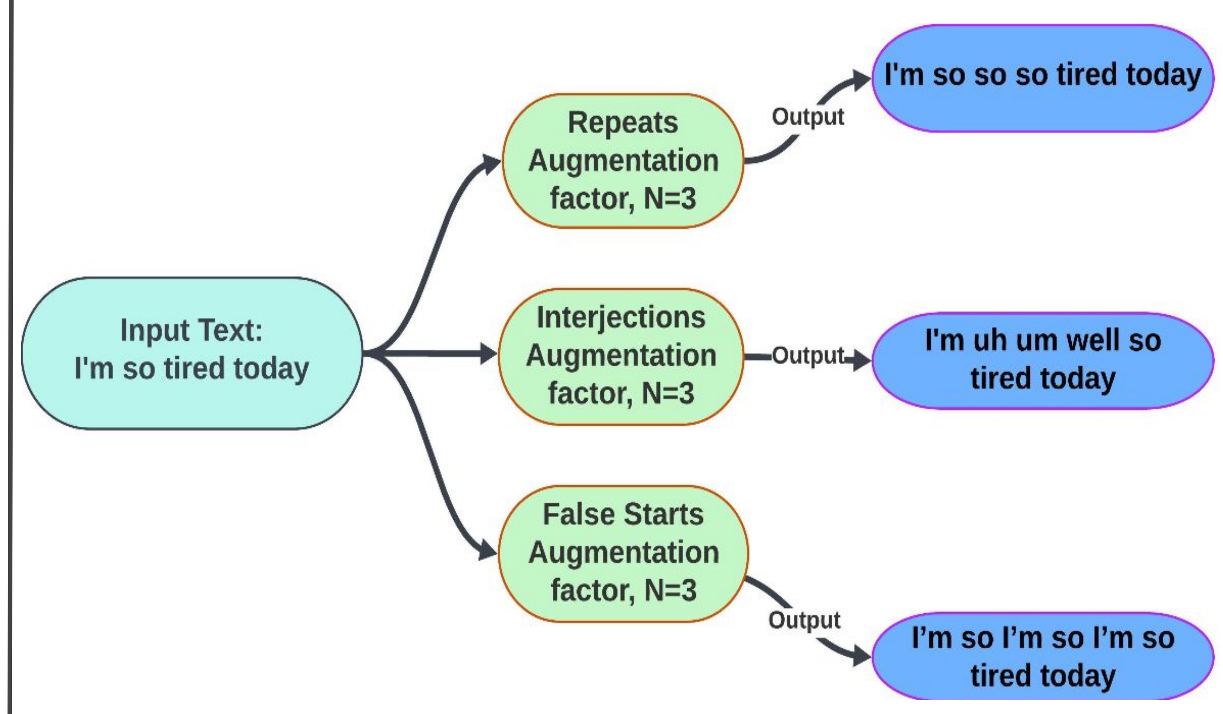
Texas A&M University

Motivation

- Disfluencies are common in everyday speech [1].
- Data which does not contain disfluencies is beneficial for performance on downstream tasks for conversational systems, summarization, and machine translation systems [2,3,4,5,6].

Research Questions

- ***RQ1:*** *How will DACL perform with Curriculum Learning on In-Domain Datasets?*
- ***RQ2:*** *How will DACL perform with Curriculum Learning on Out-of-Domain Datasets?*



- We consider T5-base [9] as our backbone model for the training stages.
- As we perform CL we see the scores gradually increasing. The R initially is low but gradually increases showing its ability to understand disfluencies better with each step.
  - DACL-Best has the highest ROUGE scores and decent P, R, and F1 scores.
- In non-CL studies, the models exhibit low P but high R scores on the SWB test set.
- This skip-CL study shows the importance of the intermediate steps in the CL process.
- Directly fine-tuning T5-base on SWB yields decent scores. However, the model returns empty strings in few cases.
- Fine-tuning DACL-Best on SWB yields a model that exhibits the highest P and shows decent R and F scores.
- Overfitting DACL-Best makes it identify more words and phrases as disfluencies at the cost of P.

RQ1: DACL on In-Domain (Spotify)

- We perform Disfluency Augmentations on 1,020 podcasts from the Spotify Podcasts Dataset [7].
  - **Repeat and Interjection augmentations**: Samples are drawn from *X~N (μ=10, σ=1)* for finding the positions where we repeat the last word or inject interjections N times; interjections are randomly selected from: *uh, um, well, like, so, okay, you know, I mean.*
  - **False start augmentations**, sentences with at least 4 words are sampled with 80% probability, and the first 2 words of the selected sentences are injected N times.
- This dataset is in-domain for the **Switchboard (SWB)** [8] dataset, as both are transcribed spoken text.

Method	**Word-Based**			**ROUGE**		
	P	**R**	**F**	**1**	**2**	**L**
Repeats Augmented $[0, 1, 5, 10]$ shuffled (no CL)	26.35	46.99	33.76	73.47	68.43	73.33
Interjections Augmented $[0, 1, 5, 10]$ shuffled (no CL)	27.69	**48.35**	**35.22**	70.33	66.52	70.18
Repeats $0-0$	69.69	0.43	0.85	89.09	83.43	89.10
Repeats $0-0, 1-0$	93.87	8.72	15.96	89.64	83.97	89.65
Repeats $0-0, 1-0, 5-0$	95.72	9.80	17.78	89.77	84.09	89.78
Repeats $0-0, 1-0, 5-0, 10-0$	**95.76**	10.04	18.18	89.79	84.11	89.78
Repeats $0-0, 10-0$	92.09	4.29	8.21	89.32	83.65	89.32
(DACL-Best) Repeats $0-0, 1-0, 5-0, 10-0$, Interjections $10-0$, False Starts $10-0$	94.80	14.74	25.52	**90.14**	**84.62**	**90.13**

Curriculum Learning on Spotify	**Fine-tune on SWB?**	**Word-Based**			**ROUGE**		
		P	**R**	**F**	**1**	**2**	**L**
No (T5-base)	**N**	17.74	49.25	26.08	0.5722	0.5124	0.5696
	Y	93.57	83.66	88.34	0.9752	0.9598	0.9750
DACL-Best	**N**	94.80	14.74	25.52	0.9015	0.8463	0.9014
	Y, 14 epochs – DACL+FT	**97.10**	84.75	90.50	0.9795	0.9650	0.9793
	Y, Overfitting, 66 epochs – DACL+FT (Overfit)	96.10	**90.25**	**93.08**	**0.9855**	**0.9758**	**0.9854**

Method	**Word-based**		
	P	**R**	**F**
DACL+FT	**97.1**	84.7	90.5
DACL+FT (Overfit)	96.1	90.2	93.0
EGBC (Bach and Huang, 2019)	95.9	86.3	90.9
EGBC + residual (Bach and Huang, 2019)	96.1	86.9	91.2
Self-Trained BERT-Based Parser (ensemble) (Jamshid Lou and Johnson, 2020b)	92.5	**97.2**	**94.8**
Self-Trained BERT-Based Parser (single) (Jamshid Lou and Johnson, 2020b)	92.2	96.6	94.3
Noisy BiLSTM (Bach and Huang, 2019)	94.7	89.8	92.2
Weight sharing (Wang et al., 2018)	92.1	90.2	91.1
BiLSTM (Zayats et al., 2016)	91.6	80.3	85.9
Semi-CRF (Zayats et al., 2016)	90.0	81.2	85.4

RQ2: DACL on Out-of-Domain (WikiSplit)

- CL on Spotify outperforms WikiSplit [10] as the presence of inherent minimal speech disfluencies in the transcribed Spotify texts adds some noise to the training process making the model more capable in identifying disfluencies.

Curriculum Learning on WikiSplit	**Fine-tune on SWB?**	**Word-Based**			**ROUGE**		
		P	**R**	**F**	**1**	**2**	**L**
No (T5-base)	**N**	17.74	49.25	26.08	0.5722	0.5124	0.5696
	Y	93.57	83.66	88.34	0.9752	0.9598	0.9750
DACL-Best	**N**	71.09	68.12	69.58	0.9391	0.9086	0.9386
	Y	**95.13**	**87.00**	**90.89**	**0.9816**	**0.9691**	**0.9815**

We find that performing DACL on our in-domain dataset (Spotify) results in the best precision and favorable recall and F1 scores for the disfluency removal task.

References

[1] Elizabeth Ellen Shriberg. 1994. Preliminaries to a theory of speech disfluencies. Ph.D. thesis
[2] Sharath Rao, et al. 2007. Improving spoken language translation by automatic disfluency removal: evidence from conversational speech transcripts. In Proceedings of Machine Translation Summit XI: Papers.
[3] Eunah Cho, et al. 2014. Tight integration of speech disfluency removal into SMT. In Conference of the European Chapter of the Association for Computational Linguistics, pages 43–47.
[4] Shaolei Wang, et al. 2020. Multi-task self-supervised learning for disfluency detection. In AAAI Conference on Artificial Intelligence, volume 34, pages 9193–9200.
[5] Hany Hassan, et al. 2014. Segmentation and disfluency removal for conversational speech translation. In Interspeech, pages 318–322.
[6] Maria Teleki, et al. 2024. Quantifying the impact of disfluency on spoken content summarization. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation.
[7] Clifton, et al. 2020. 100,000 podcasts: A spoken English document corpus.
[8] Godfrey, John J and Holliman, Edward. 1997. Switchboard-1 Release 2.
[9] Colin Raffel, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 21(1):5485–5551.
[10] Jan Botha, et al. 2018. Learning To Split and Rephrase From Wikipedia Edit History.

Texas A&M University
Department of Computer Science & Engineering

Code ***R. Chaudhury***