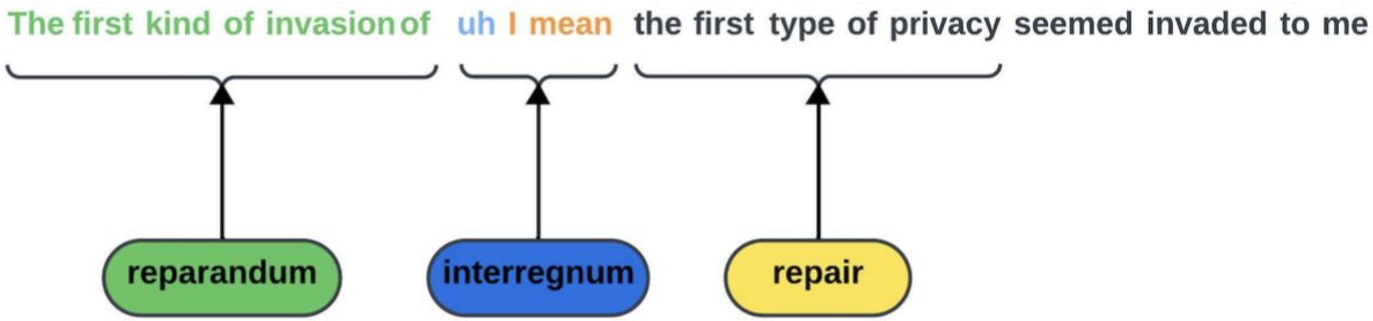
DACL: Disfluency Augmented Curriculum Learning for Fluent Text Generation

Rohan Chaudhury, Maria Teleki, Xiangjue Dong, James Caverlee
Texas A&M University
College Station, TX, USA
{rohan.chaudhury, mariateleki, xj.dong, caverlee}@tamu.edu

In LREC-COLING 2024

1 Minute Summary


The first kind of invasion of
uh
I mean
the first type of privacy seemed invaded to me
reparandum
interregnum
repair

- In this work we aim at disfluency removal from transcribed spoken text.
- **What are disfluencies?**
  - Disfluencies are any breaks in the regular flow of speech, such as false starts, corrections, repeats, and filled pauses.
- **Motivation:**
  - We prefer having less false positives (higher Precision) than less false negatives (higher Recall). That is, we do not want to remove important information at the cost of removing most disfluencies in the text.

1 Minute Summary

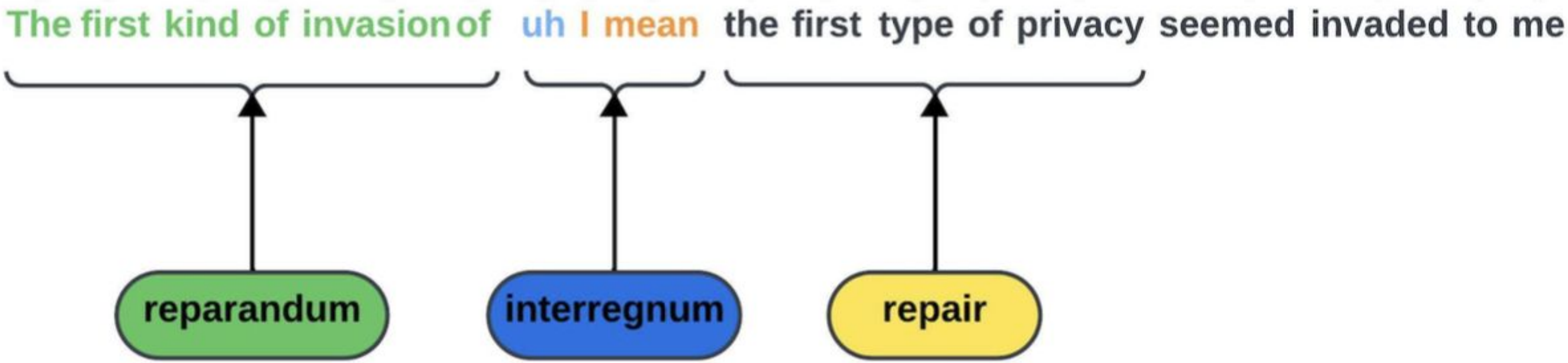
- We propose **Disfluency Augmented Curriculum Learning (DACL)** approach to tackle the complex structure of disfluent sentences and generate fluent texts from them.
  - We first **synthetically create texts with various levels of disfluency** from clean texts.
  - DACL harnesses the tiered structure of our generated synthetic disfluent data using **Curriculum Learning**, by training the model on basic samples (i.e. more fluent) first before training it on more complex samples (i.e. more disfluent).
- **Our model surpasses existing techniques** in word-based precision (by up to 1%) and has shown favorable recall and F1 scores on the widely used Switchboard Penn Treebank-3 dataset.

DACL: Disfluency Augmented Curriculum Learning for Fluent Text Generation

Rohan Chaudhury, Maria Teleki, Xiangjue Dong, James Caverlee
Texas A&M University
College Station, TX, USA
{rohan.chaudhury, mariateleki, xj.dong, caverlee}@tamu.edu

In LREC-COLING 2024

PARTS OF SPEECH DISFLUENCY [SHRIBERG Ph.D. THESIS 1994]



The first kind of invasion of uh I mean the first type of privacy seemed invaded to me

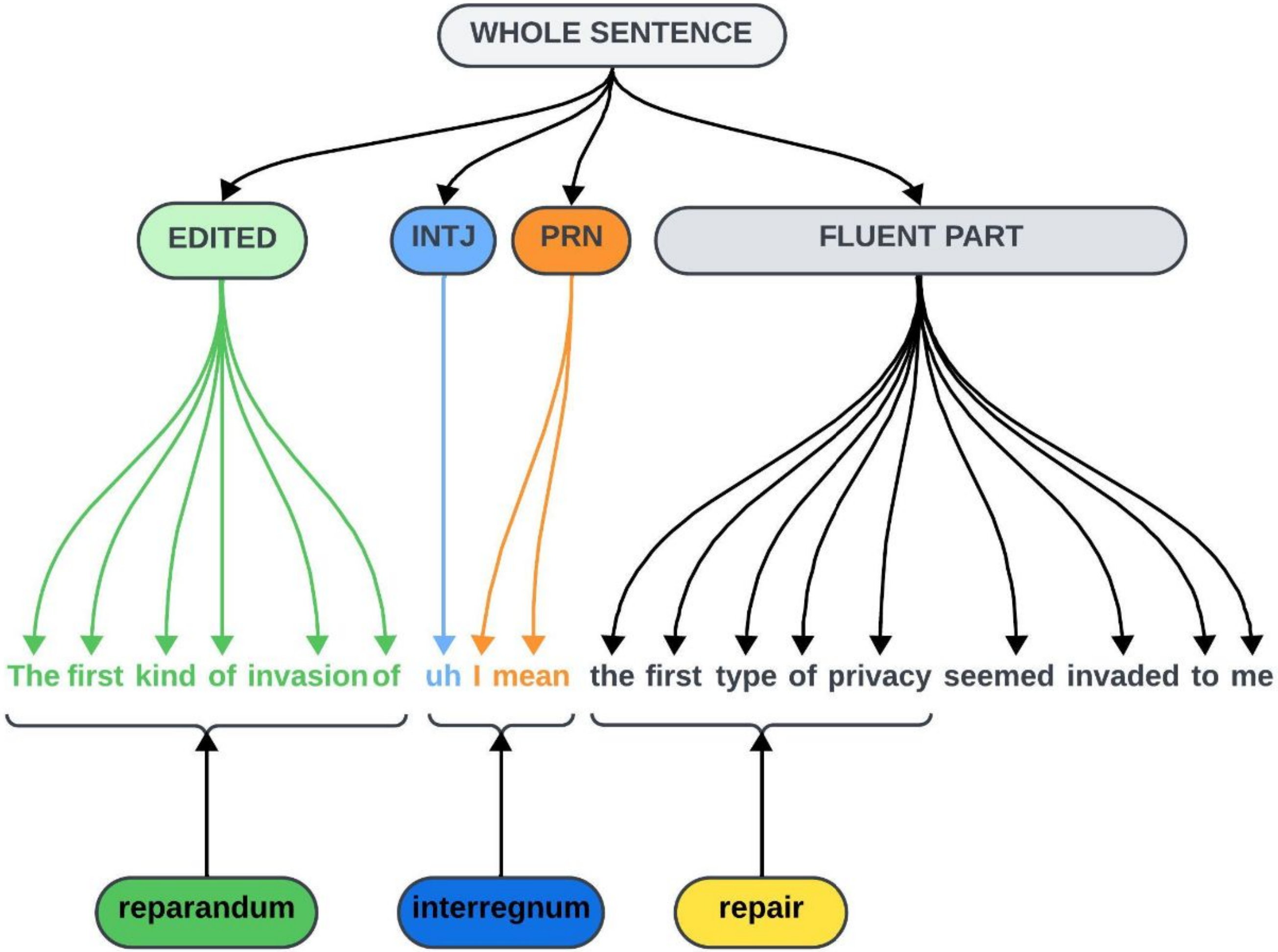
reparandum

interregnum

repair

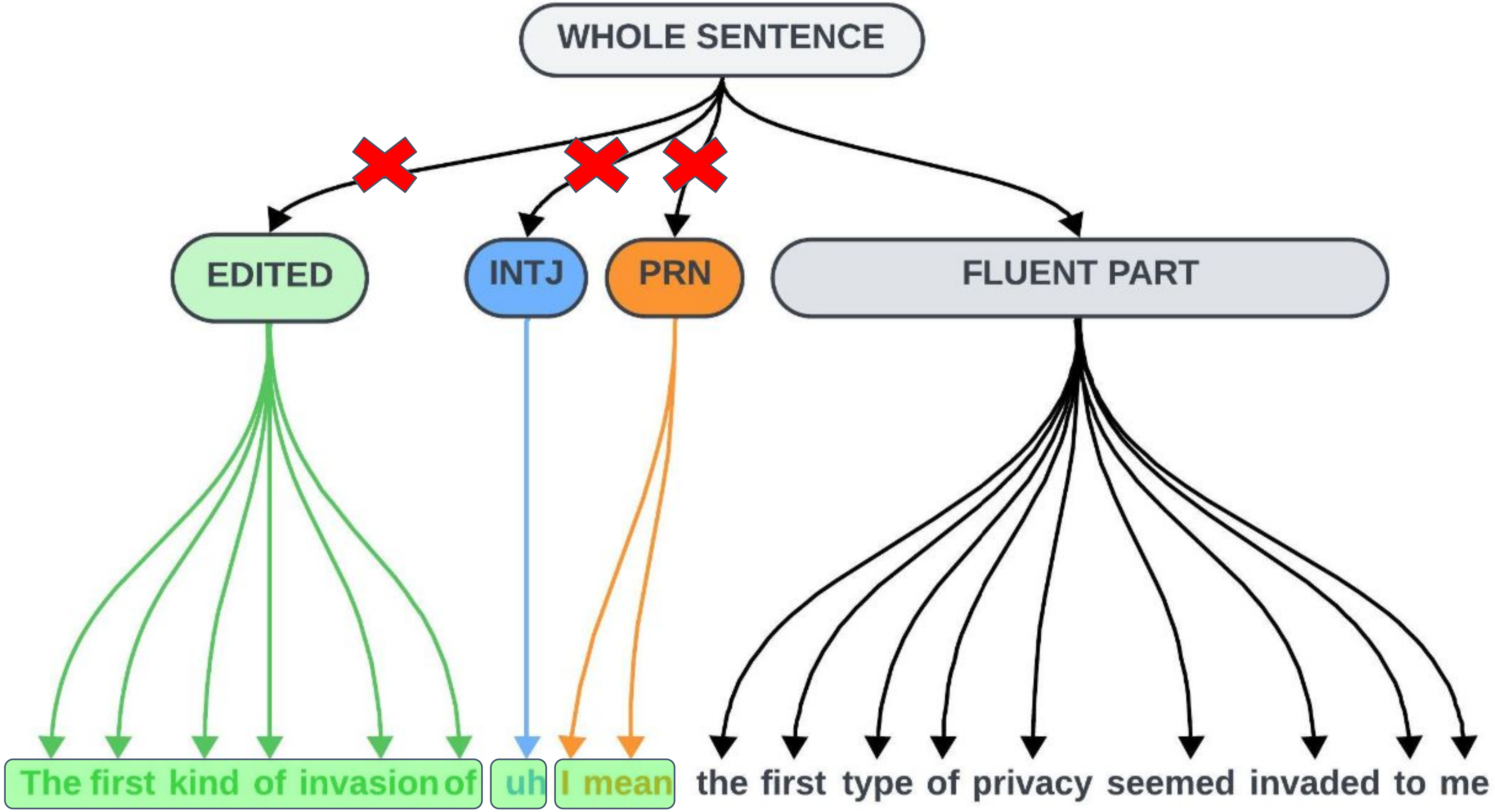
E. E. Shriberg, Preliminaries to a theory of speech disfluencies. PhD thesis, Citeseer, 1994

MAJOR KINDS OF DISFLUENCY TAGS IN SWB [LOU ET AL. ACL 2020]


WHOLE SENTENCE
EDITED
INTJ
PRN
FLUENT PART
The first kind of invasion of
uh
I mean
the first type of privacy seemed invaded to me
reparandum
interregnum
repair

Lou et al., “Improving disfluency detection by self-training a self- attentive model,” in Association for Computational Linguistics, July 2020.

HOW THE FLUENT TEXT WOULD LOOK LIKE HERE:


WHOLE SENTENCE
EDITED
INTJ
PRN
FLUENT PART
The first kind of invasion of
uh
I mean
the first type of privacy seemed invaded to me

In the task of Disfluency Removal we would prefer having less **false positives** than less **false negatives.** That is, we do not want to remove important information at the cost of removing most the disfluencies in the text.

Which also translates to us wanting more **Precision** than more **Recall**

EXPERIMENTAL SETUP

- **T5-Base** is our backbone model for all the experiments (developed by Google AI)
- Two types of dataset:
  - **Pre-training Datasets**:
    - Spotify Podcast Dataset - in-domain dataset
    - The WikiSplit Dataset - out-of-domain dataset
  - **Evaluation Dataset**:
    - Switchboard, Treebank-3

- Comparing **word-based Precision, Recall ,F1 scores** and **ROUGE** scores of the models
- Motivation is to **increase** the **word-based Precision scores** while also maintaining favorable Recall, F1, and ROUGE scores.

Raffel et al., “Exploring the limits of transfer learning with a unified text-to-text transformer,” in JMLR 2020

RESEARCH QUESTIONS

RQ1: How will DACL perform with Curriculum Learning on in-domain datasets?

- ❖ We build a synthetic disfluency augmentation approach that produces progressive versions of the original in-domain texts (transcribed speech) with various degrees of disfluencies. We then use Curriculum Learning using the augmented dataset to train a T5-base model for the task of disfluency removal from the Switchboard dataset .

RQ2: How will DACL perform with Curriculum Learning on out-of-domain datasets?

- ❖ We employ the same methodology as in RQ1 for RQ2, substituting in-domain texts with out-of-domain texts (written text) to create the synthetically augmented dataset for the CL process.

RQ1: INTUITION

- We train a **sequence to sequence generation** model incrementally on increasing levels of **augmented disfluencies**

- This process is known as **Curriculum Learning**

- This increases the models ability to **better identify disfluencies from normal fluent text** by facilitating gradual learning and incrementally increasing the generalization capability and robustness of the model to disfluencies

RQ1: SYNTHETIC DISFLUENCY AUGMENTATION
[WANG ET AL. AAAI 2020, PASSALI ET AL. LREC 2022]

We do 3 kinds of **synthetic disfluency augmentations**:

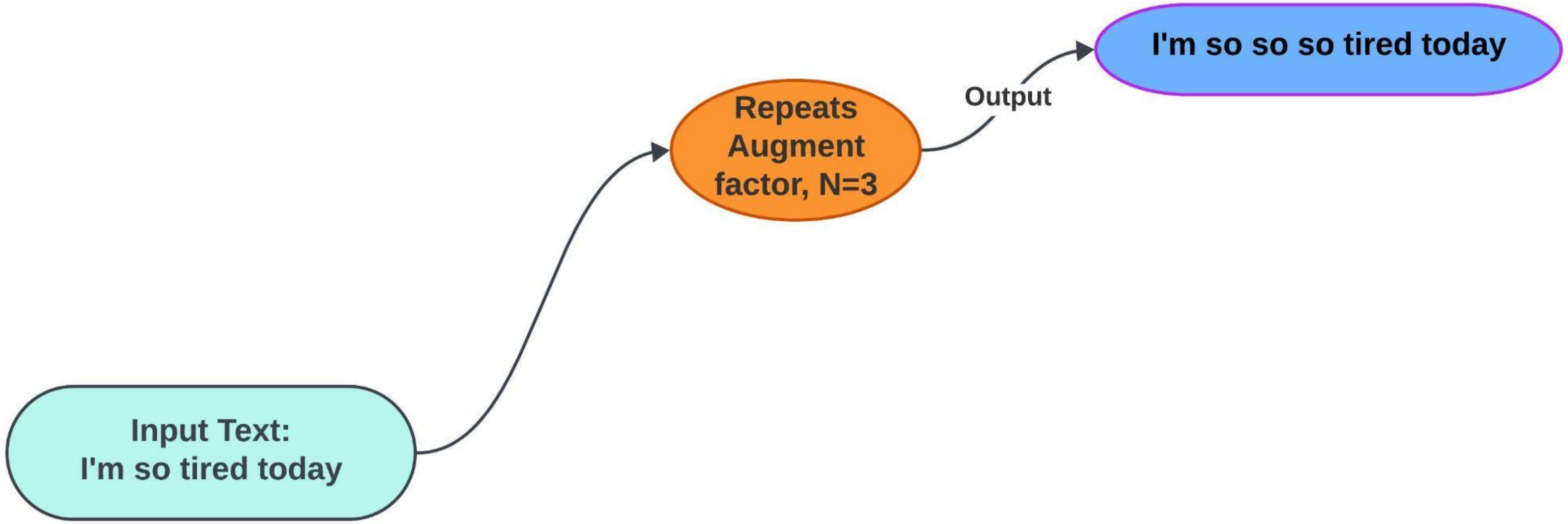
❑ **Repeats:**
Repeated words

❑ **Interjection:**
“uh”, “um”, “well”, “like”, “so”, “okay”, “you know”, “I mean”

❑ **False-Starts:**
Edited phrases

Wang et al., “Multi-task self-supervised learning for disfluency detection,” in Proceedings of the AAAI Conference on Artificial Intelligence, July 2020.
Passali et al., “LARD: Large- scale artificial disfluency generation,” in Language Resources and Evaluation Conference, June 2022

RQ1: SYNTHETIC DISFLUENCY AUGMENTATION

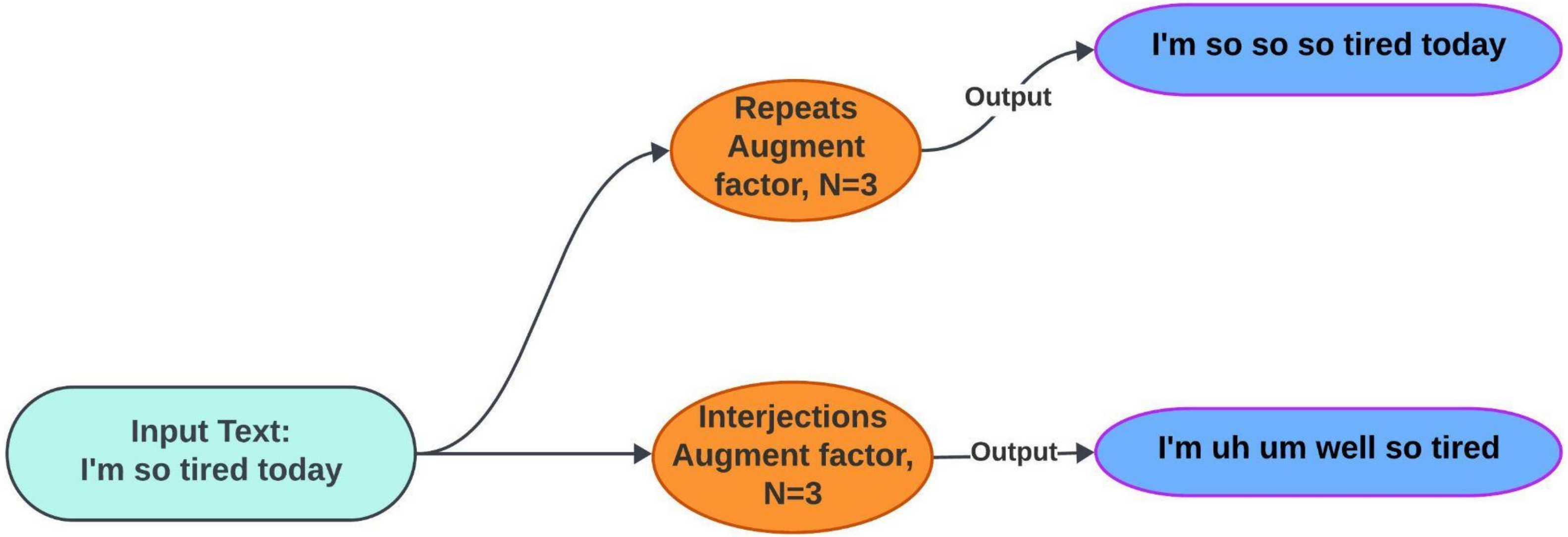


Input Text:
I'm so tired today

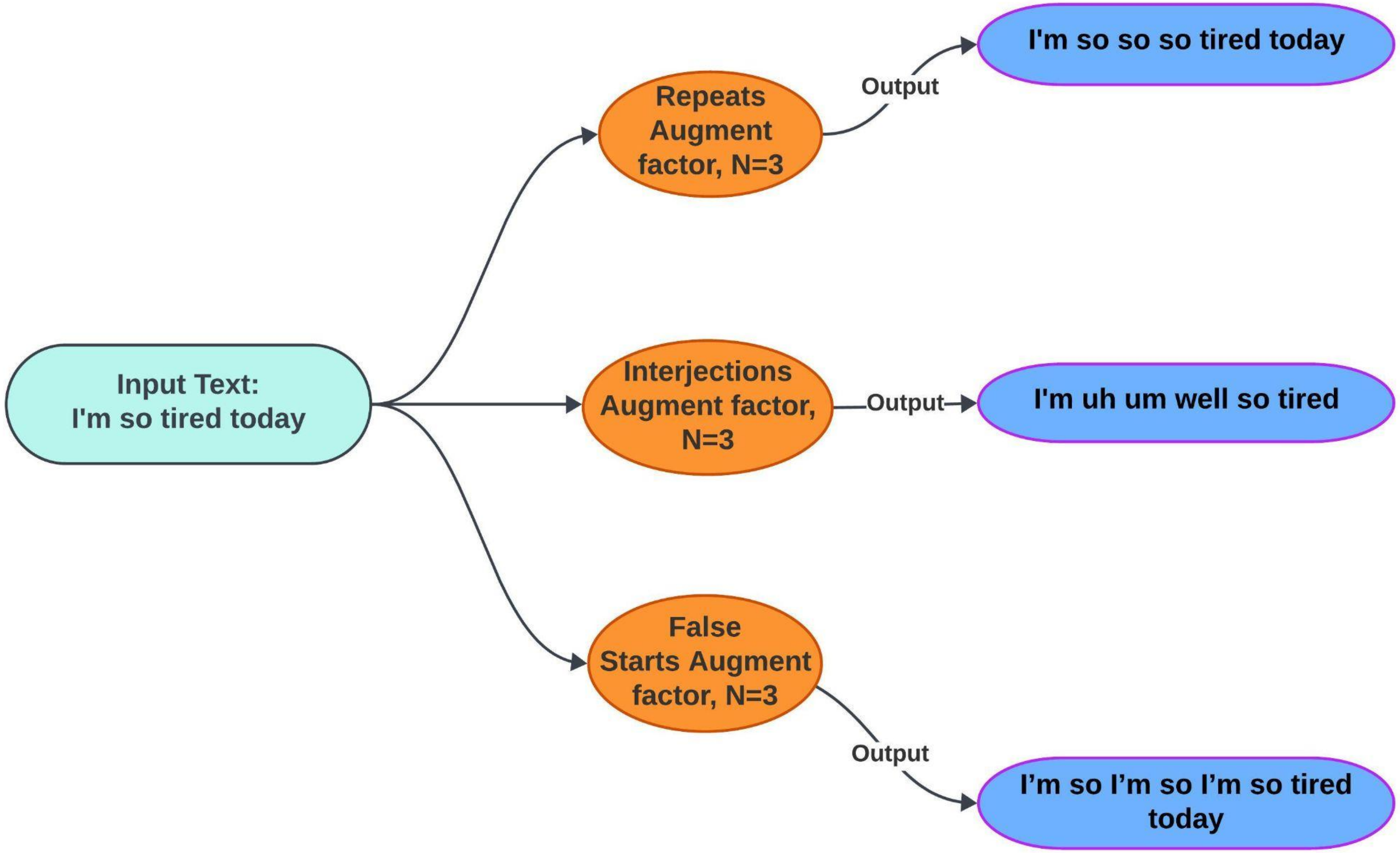
Repeats Augment factor, N=3

I'm so so so tired today

RQ1: SYNTHETIC DISFLUENCY AUGMENTATION



RQ1: SYNTHETIC DISFLUENCY AUGMENTATION



Input Text:
I'm so tired today

Repeats Augment factor, N=3

Interjections Augment factor, N=3

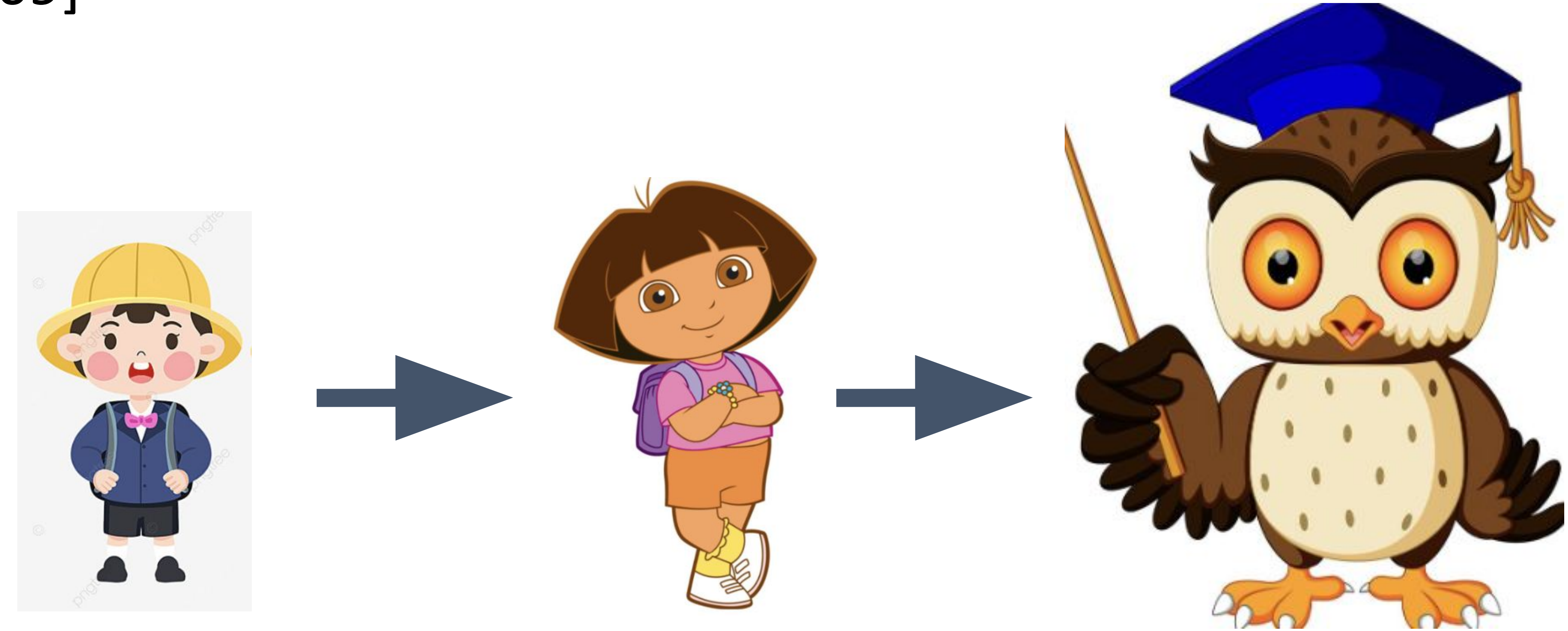
False Starts Augment factor, N=3

I'm so so so tired today

I'm uh um well so tired

I’m so I’m so I’m so tired today

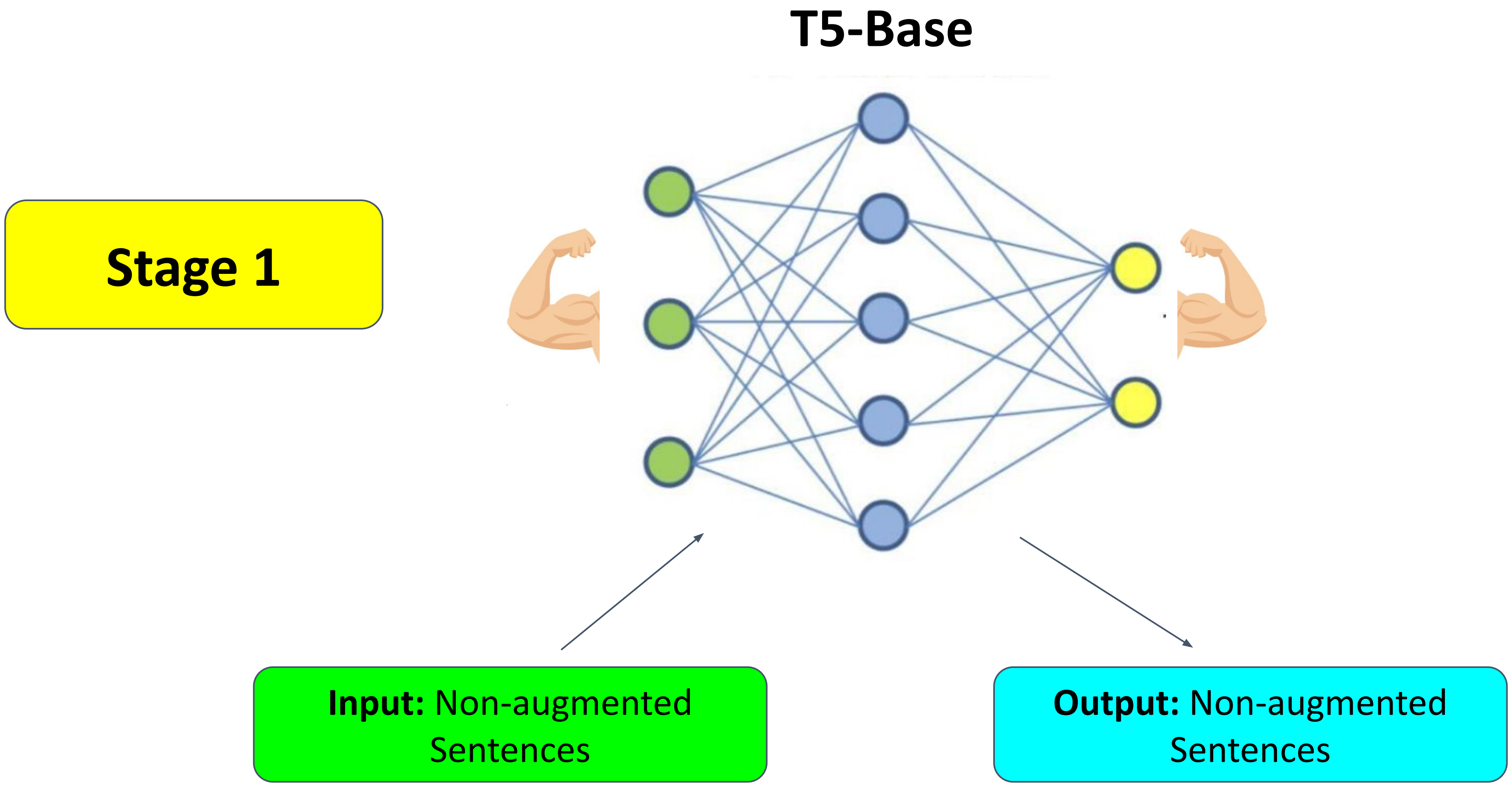
RQ1: CURRICULUM LEARNING [BENGIO ET AL. ICML 2009]



Innocent Pre-schooler **Highschool Explorer** **Night Owl Graduate**

Bengio et al., “Curriculum learning,” in Proceedings of the 26th annual international conference on machine learning, pp. 41–48, 2009.

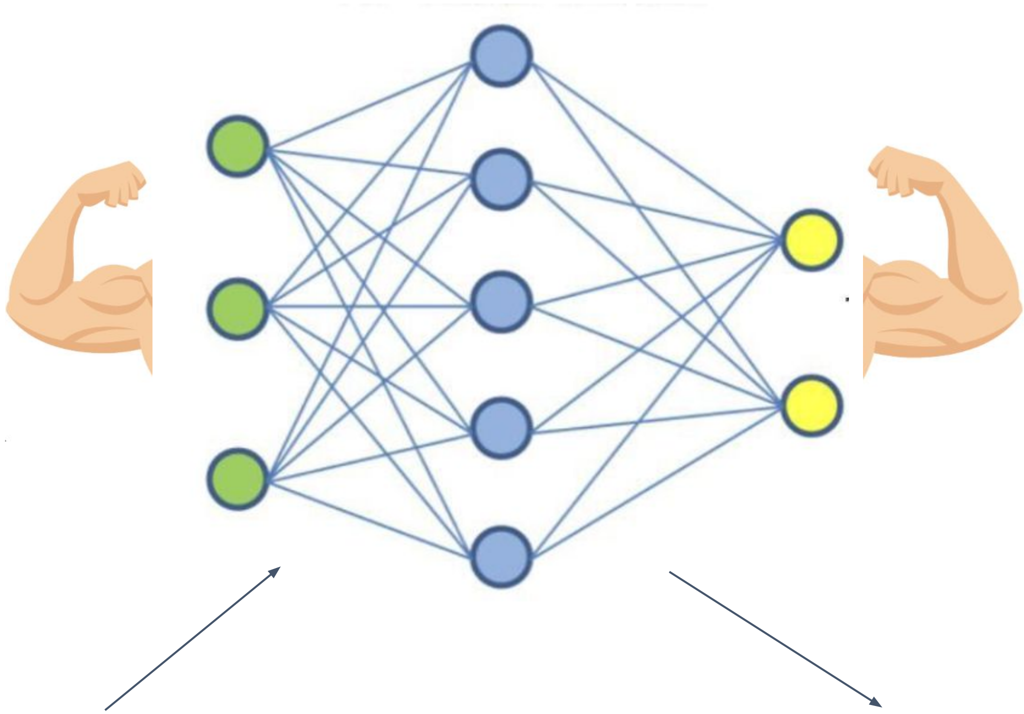
RQ1: CURRICULUM LEARNING PROCESS


T5-Base
Stage 1
Input: Non-augmented Sentences
Output: Non-augmented Sentences

Raffel et al., “Exploring the limits of transfer learning with a unified text-to-text transformer,” in JMLR 2020

RQ1: CURRICULUM LEARNING PROCESS

Stage 2

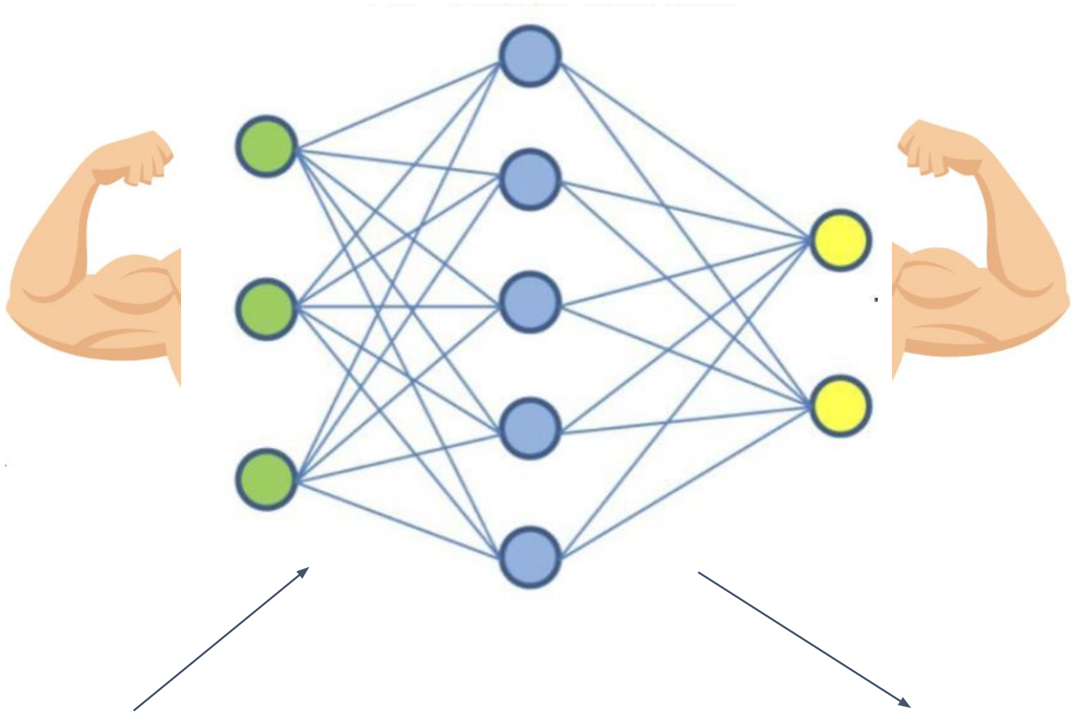


Input: Repeat augmented Sentences **(Level 1)**

Output: Non-augmented Sentences

RQ1: CURRICULUM LEARNING PROCESS

Stage 3

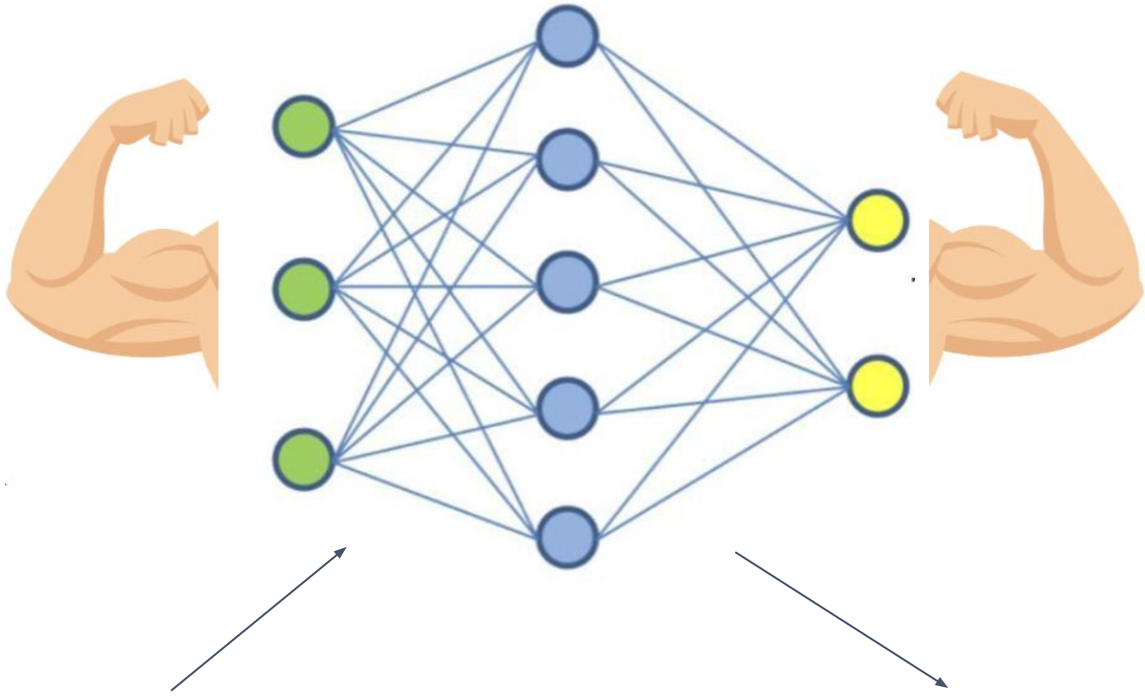


Input: Repeat augmented Sentences **(Level 5)**

Output: Non-augmented Sentences

RQ1: CURRICULUM LEARNING PROCESS

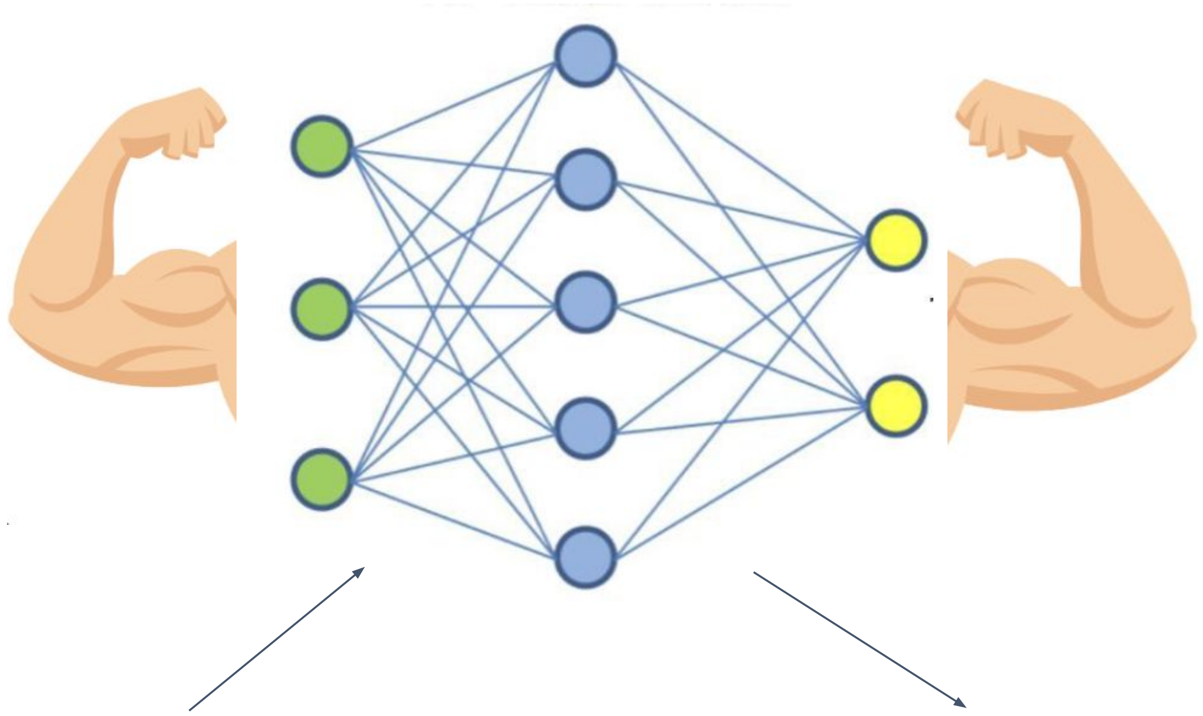
Stage 4



Input: Repeat augmented Sentences **(Level 10)**

Output: Non-augmented Sentences

RQ1: CURRICULUM LEARNING PROCESS



Stage 5

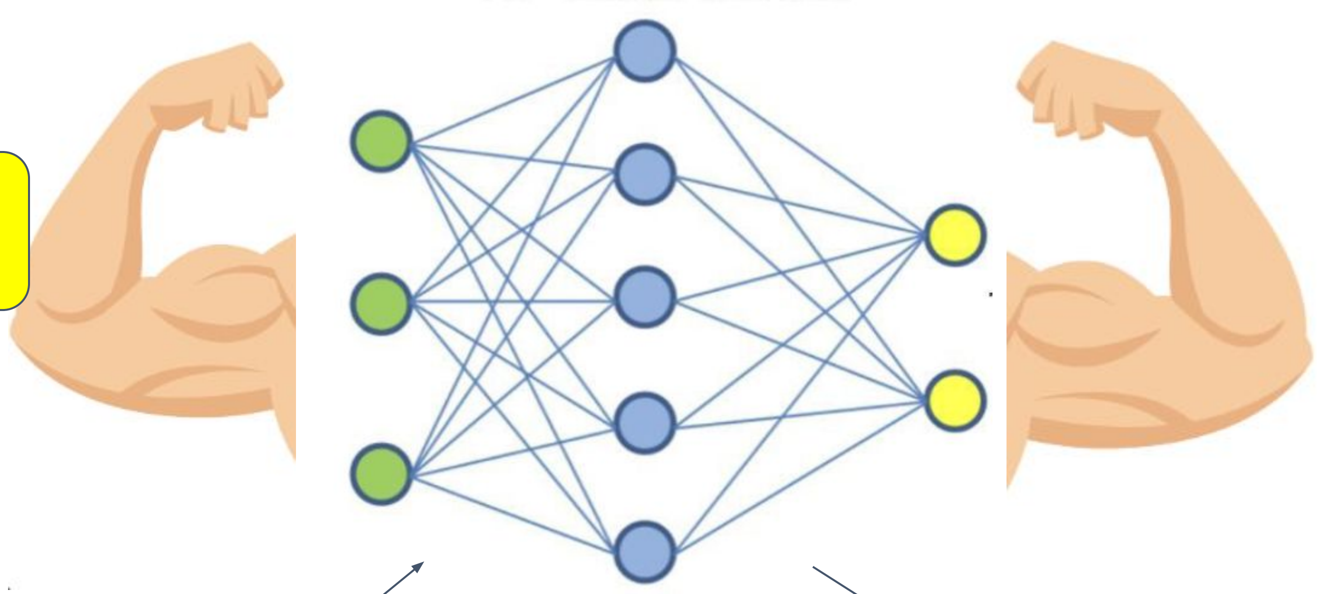
Input: Interjection augmented Sentences **(Level 10)**

Output: Non-augmented Sentences

RQ1: CURRICULUM LEARNING PROCESS

CL-Best

Stage 6



Number of **model parameters** did not increase, only its **capability** increased

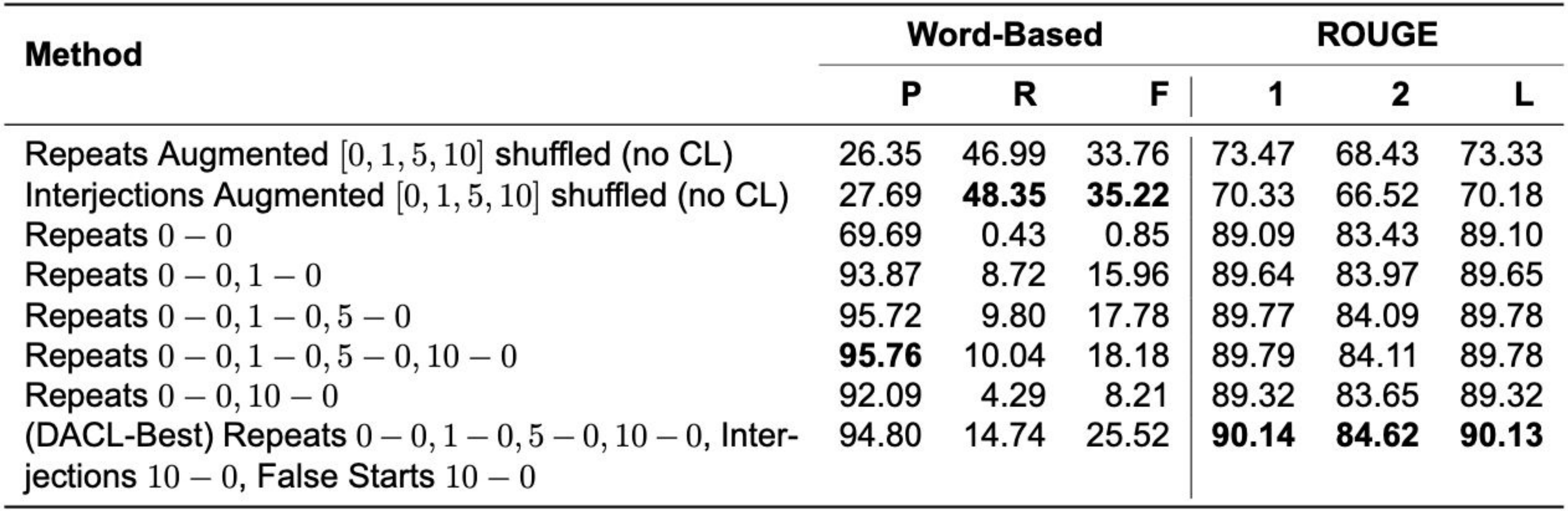
Input: False-Start augmented Sentences **(Level 10)**

Output: Non-augmented Sentences

RQ1: FINE-TUNING ON TARGET DATASET

- ❏ Through the DACL process we obtain the DACL-Best model that understands the basics of what is a disfluency and can precisely remove disfluencies that we have shown through our synthetic augmentations.
- ❏ However, naturally disfluent datasets like Switchboard Treebank-3 can have more kinds of disfluencies that we have not engineered in our augmentations.
- ❏ To tackle these new types of disfluencies, we fine-tune DACL-best – the best checkpoint from DACL – on the training set of the Evaluation dataset, Switchboard Treebank-3.

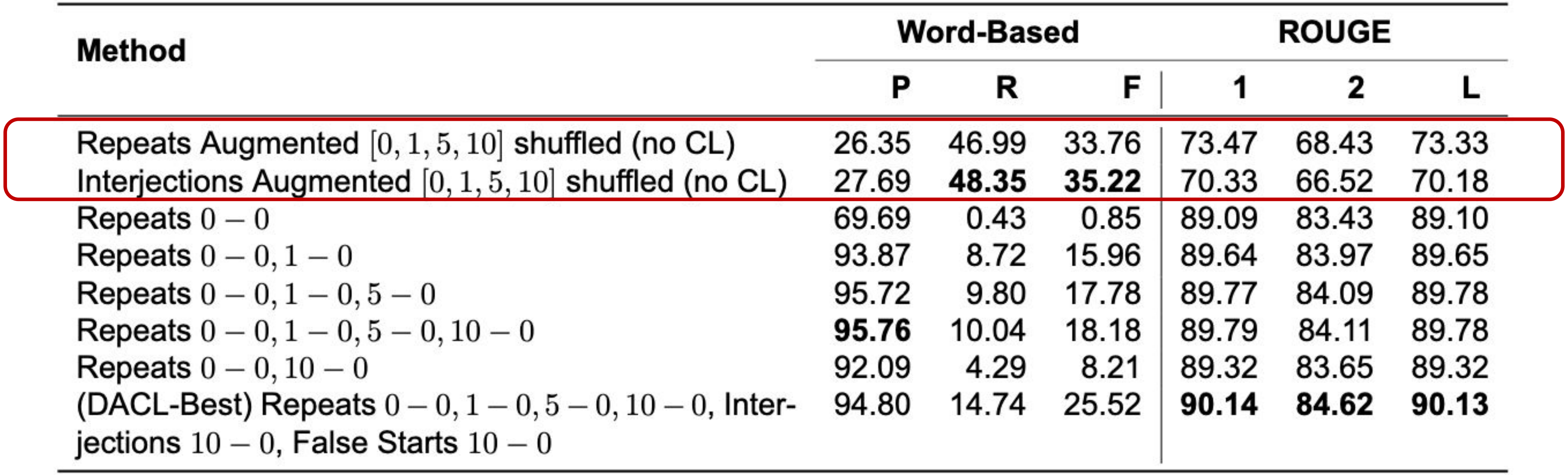
RQ1: DACL ON SPOTIFY: HOW IT HELPS?



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**

The table shows the scores after evaluating the respective models on the Switchboard test set

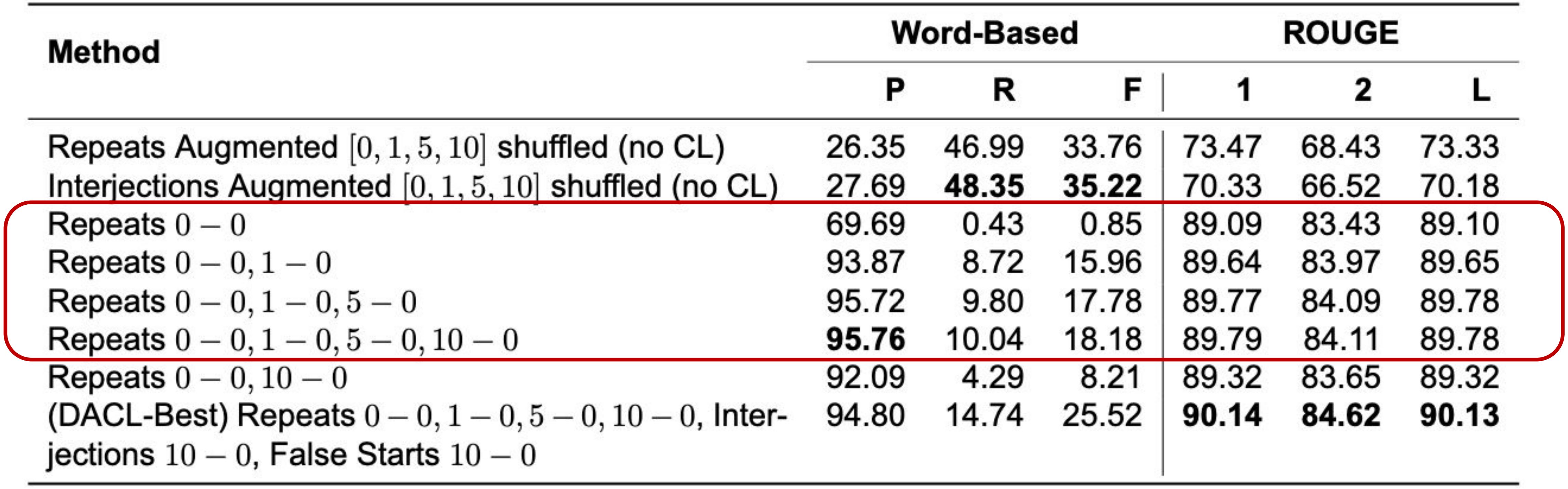
RQ1: DACL ON SPOTIFY: HOW IT HELPS?



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**

These models exhibit low precision but high recall scores as they frequently return empty strings and random tokens

RQ1: DACL ON SPOTIFY: HOW IT HELPS?

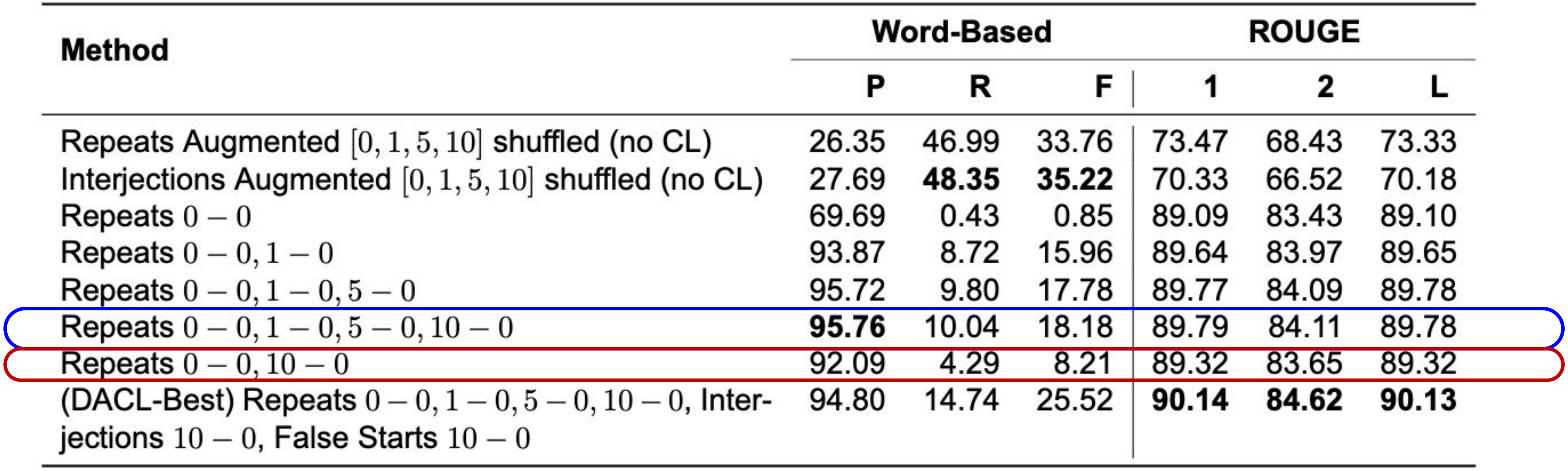


Method	Word-Based			ROUGE		
	P	R	F	1	2	L
Repeats Augmented $[0, 1, 5, 10]$ shuffled (no CL)	26.35	46.99	33.76	73.47	68.43	73.33
Interjections Augmented $[0, 1, 5, 10]$ shuffled (no CL)	27.69	**48.35**	**35.22**	70.33	66.52	70.18
Repeats $0-0$	69.69	0.43	0.85	89.09	83.43	89.10
Repeats $0-0, 1-0$	93.87	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**

As we perform CL we see the scores gradually increasing.

The recall initially is low but gradually increases showing its ability to understand disfluencies better with each step.

RQ1: DACL ON SPOTIFY: HOW IT HELPS?

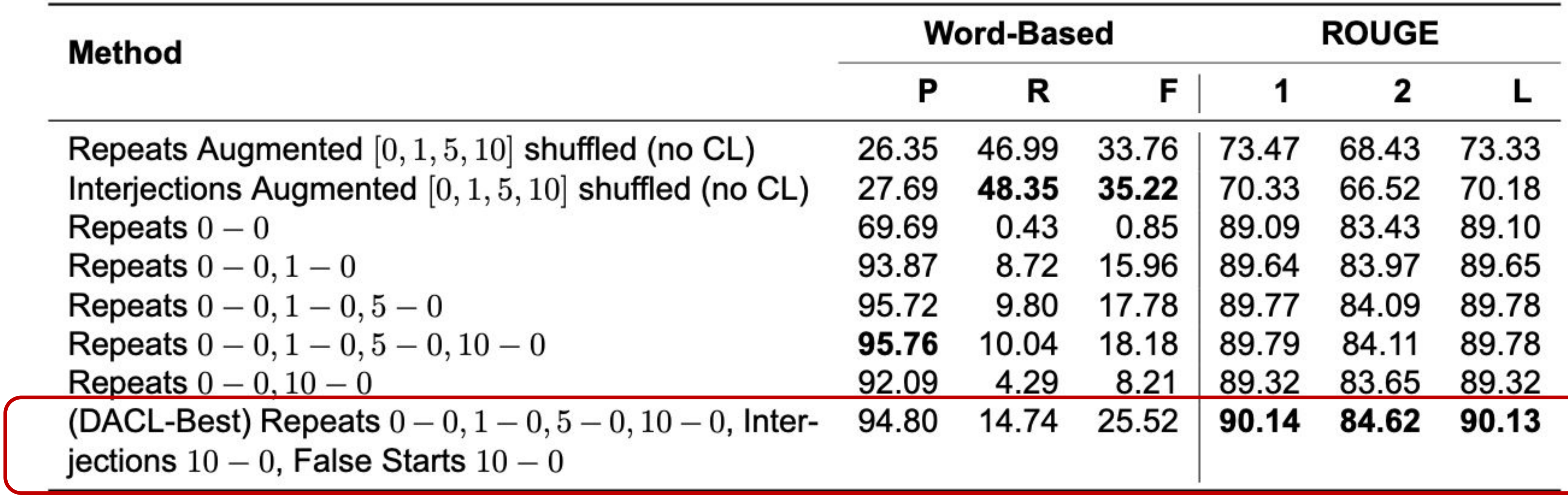


Method	Word-Based			ROUGE		
	P	R	F	1	2	L
Repeats Augmented $[0, 1, 5, 10]$ shuffled (no CL)	26.35	46.99	33.76	73.47	68.43	73.33
Interjections Augmented $[0, 1, 5, 10]$ shuffled (no CL)	27.69	**48.35**	**35.22**	70.33	66.52	70.18
Repeats $0 - 0$	69.69	0.43	0.85	89.09	83.43	89.10
Repeats $0 - 0, 1 - 0$	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**

This ablation study shows the importance of the intermediate steps in the Curriculum learning process.

The scores from the row outlined in blue that has 2 additional steps of CL

RQ1: DACL ON SPOTIFY: HOW IT HELPS?

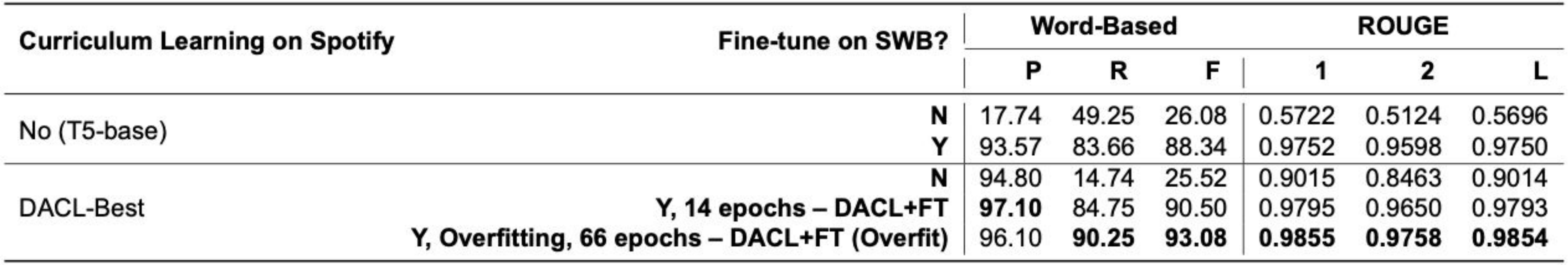


Method	Word-Based			ROUGE		
	P	R	F	1	2	L
Repeats Augmented $[0, 1, 5, 10]$ shuffled (no CL)	26.35	46.99	33.76	73.47	68.43	73.33
Interjections Augmented $[0, 1, 5, 10]$ shuffled (no CL)	27.69	**48.35**	**35.22**	70.33	66.52	70.18
Repeats $0-0$	69.69	0.43	0.85	89.09	83.43	89.10
Repeats $0-0, 1-0$	93.87	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**

The model has the highest ROUGE scores and decent P, R, F1 scores

This indicates that the CL process is instrumental in increasing the quality of the generated outputs

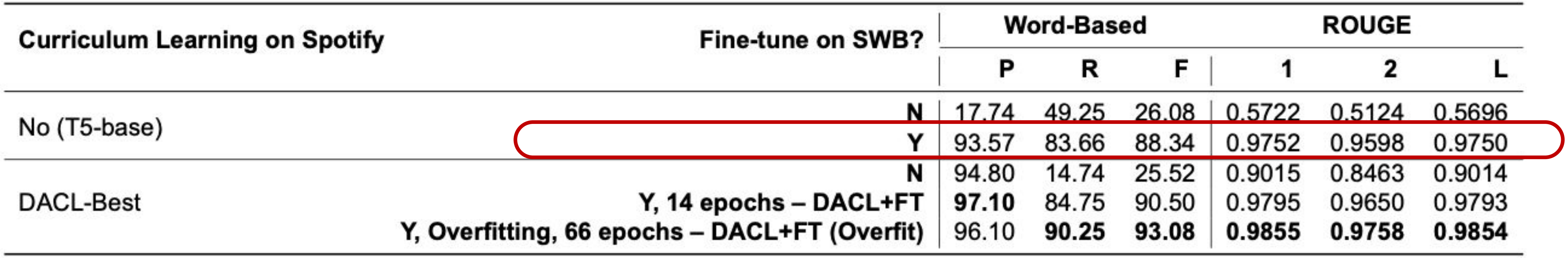
RQ1: DACL ON SPOTIFY + FINE-TUNING ON SWITCHBOARD



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**

This table shows the results of fine-tuning on Switchboard training set and evaluating on Switchboard test set

RQ1: DACL ON SPOTIFY + FINE-TUNING ON SWITCHBOARD

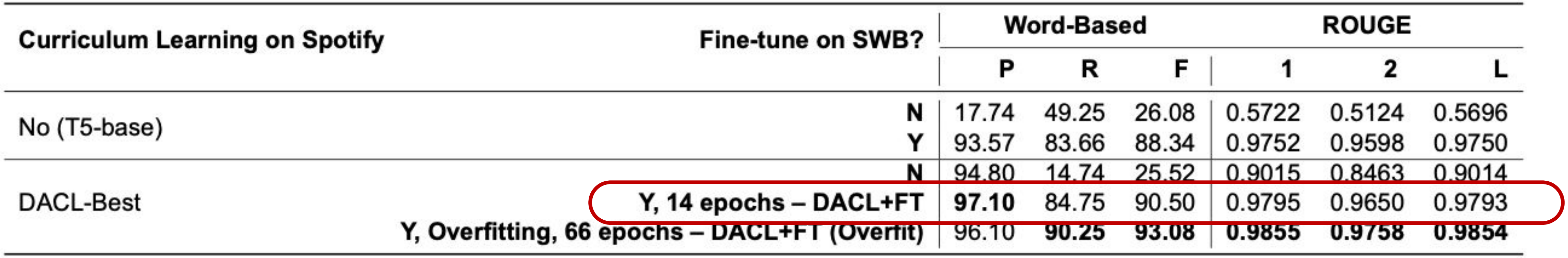


Curriculum Learning on Spotify	Fine-tune on SWB?	Word-Based			ROUGE		
		P	R	F	1	2	L
No (T5-base)	N	17.74	49.25	26.08	0.5722	0.5124	0.5696
	Y	93.57	83.66	88.34	0.9752	0.9598	0.9750
DACL-Best	N	94.80	14.74	25.52	0.9015	0.8463	0.9014
	Y, 14 epochs – DACL+FT	**97.10**	84.75	90.50	0.9795	0.9650	0.9793
	Y,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

Directly fine-tuning T5-base on Switchboard yields decent scores

However, the model suffers from returning empty strings in few cases

RQ1: DACL ON SPOTIFY + FINE-TUNING ON SWITCHBOARD

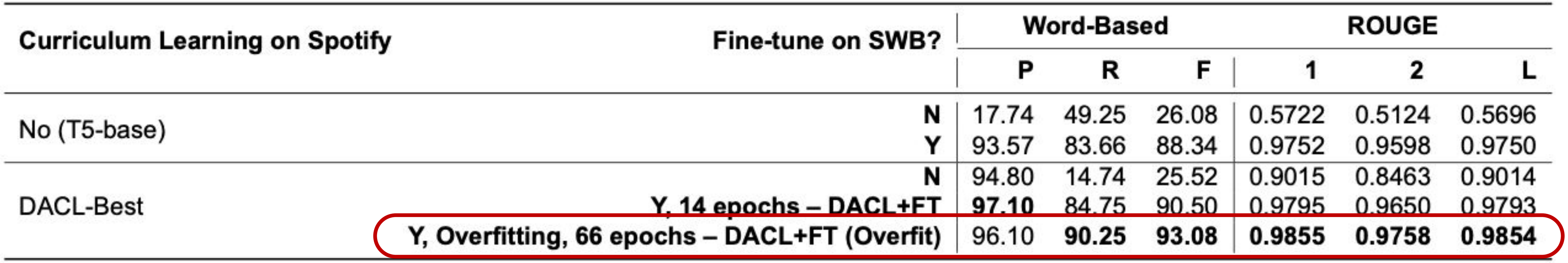


Curriculum Learning on Spotify	Fine-tune on SWB?	Word-Based			ROUGE		
		P	R	F	1	2	L
No (T5-base)	N	17.74	49.25	26.08	0.5722	0.5124	0.5696
	Y	93.57	83.66	88.34	0.9752	0.9598	0.9750
DACL-Best	N	94.80	14.74	25.52	0.9015	0.8463	0.9014
	Y, 14 epochs – DACL+FT	**97.10**	84.75	90.50	0.9795	0.9650	0.9793
	Y,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

This is the model with the best validation loss on the SWB test set and it has the highest Precision.

The training process converges in 14 epochs.

RQ1: DACL ON SPOTIFY + FINE-TUNING ON SWITCHBOARD

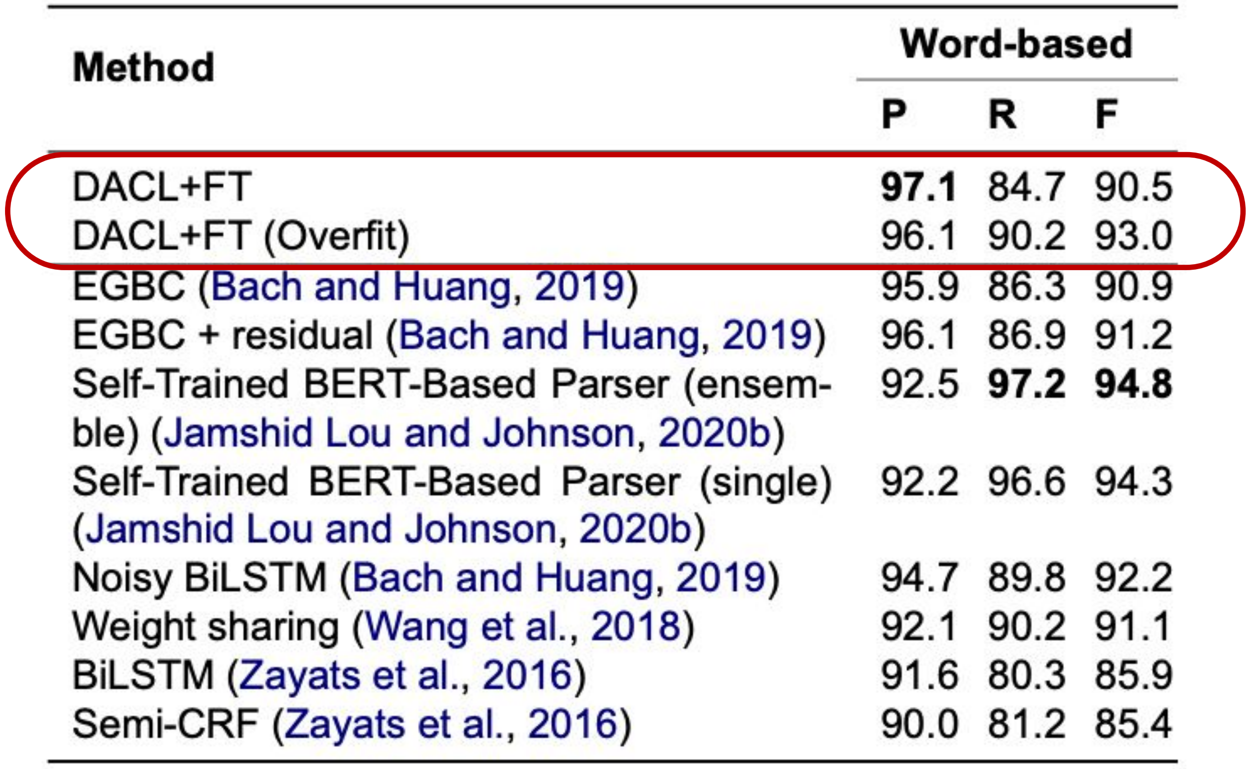


Curriculum Learning on Spotify	Fine-tune on SWB?	Word-Based			ROUGE		
		P	R	F	1	2	L
No (T5-base)	N	17.74	49.25	26.08	0.5722	0.5124	0.5696
	Y	93.57	83.66	88.34	0.9752	0.9598	0.9750
DACL-Best	N	94.80	14.74	25.52	0.9015	0.8463	0.9014
	Y, 14 epochs – DACL+FT	**97.10**	84.75	90.50	0.9795	0.9650	0.9793
	Y,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

This is the result of overfitting the model on the SWB train set till 66 epochs.

It has the highest Recall, F1, ROUGE scores.

RQ1: COMPARISON WITH PREVIOUS RESEARCH

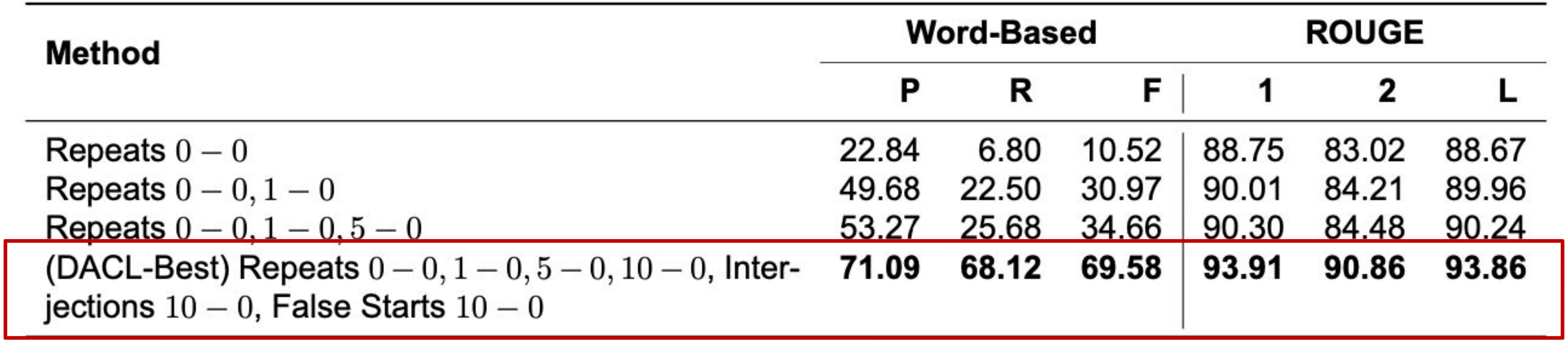


Method	Word-based		
	P	R	F
DACL+FT	**97.1**	84.7	90.5
DACL+FT (Overfit)	96.1	90.2	93.0
EGBC (Bach and Huang, 2019)	95.9	86.3	90.9
EGBC + residual (Bach and Huang, 2019)	96.1	86.9	91.2
Self-Trained BERT-Based Parser (ensemble) (Jamshid Lou and Johnson, 2020b)	92.5	**97.2**	**94.8**
Self-Trained BERT-Based 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

Table shows comparison with previous research.

Model with best validation loss has the highest precision but overfitted model has overall better scores. So why not overfit then?

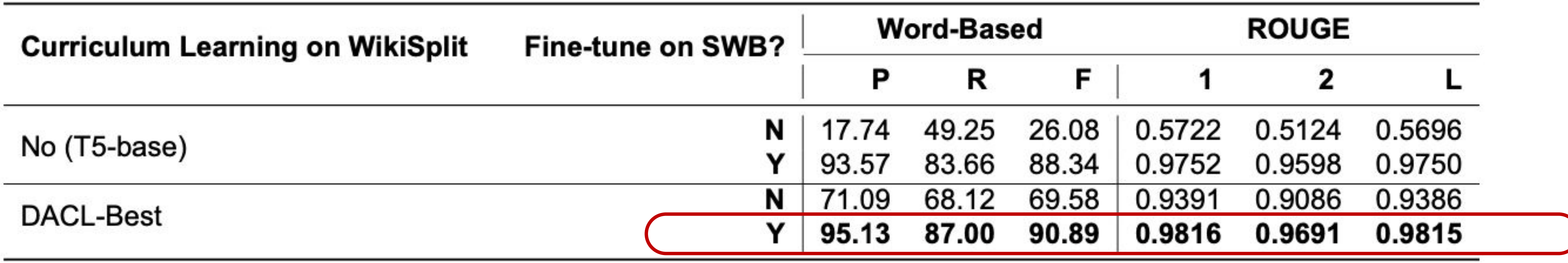
RQ2: DACL ON WIKISPLIT: WHAT IS THE DIFFERENCE?



Method	Word-Based			ROUGE		
	P	R	F	1	2	L
Repeats $0-0$	22.84	6.80	10.52	88.75	83.02	88.67
Repeats $0-0, 1-0$	49.68	22.50	30.97	90.01	84.21	89.96
Repeats $0-0, 1-0, 5-0$	53.27	25.68	34.66	90.30	84.48	90.24
(DACL-Best) Repeats $0-0, 1-0, 5-0, 10-0$, Interjections $10-0$, False Starts $10-0$	**71.09**	**68.12**	**69.58**	**93.91**	**90.86**	**93.86**

The table shows the results of fine-tuning the T5-base model on the WikiSplit training set and testing on the Switchboard test set.

RQ2: DACL ON WIKISPLIT + FINE-TUNING ON SWITCHBOARD



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**

The table shows the results of fine-tuning the WikiSplit DACL-best model on the Switchboard training set and testing on the Switchboard test set.

RQ2: WHY IS CL ON SPOTIFY GIVING BETTER PRECISION?

- This can be attributed to the fact that WikiSplit is a **written dataset** that has almost **zero speech-related disfluencies** whereas Spotify is a **spoken transcribed dataset** that has **minimal disfluencies** present in the non-augmented sentences.

- The presence of inherent minimal speech disfluencies in the Spotify dataset adds some **noise** to the entire training process and also instructs the model to better identify only the **repeats, interjections, and false start** disfluencies from the sentences and leave the rest unaltered. This makes the model more precise and judicious in its disfluency selection process.

CONCLUSION:

❖ Our experiments and ablation studies show the efficacy of our DACL process and our best model **outperforms** the state-of-the-art methods in **word-based precision** and demonstrates favorable word-based recall and F1 scores on the widely used Switchboard test set for the task of Disfluency Removal.

❖ We find that performing DACL on in-domain dataset results in the best Precision and favorable recall and F1 scores for the task of disfluency removal.

Link to our code on GitHub!

DACL: Disfluency Augmented Curriculum Learning for Fluent Text Generation

Rohan Chaudhury, Maria Teleki, Xiangjue Dong, James Caverlee
Texas A&M University
College Station, TX, USA
{rohan.chaudhury, mariateleki, xj.dong, caverlee}@tamu.edu

In LREC-COLING 2024