ACLM: A Selective-Denoising based Generative Data Augmentation Approach for Low-Resource Complex NER
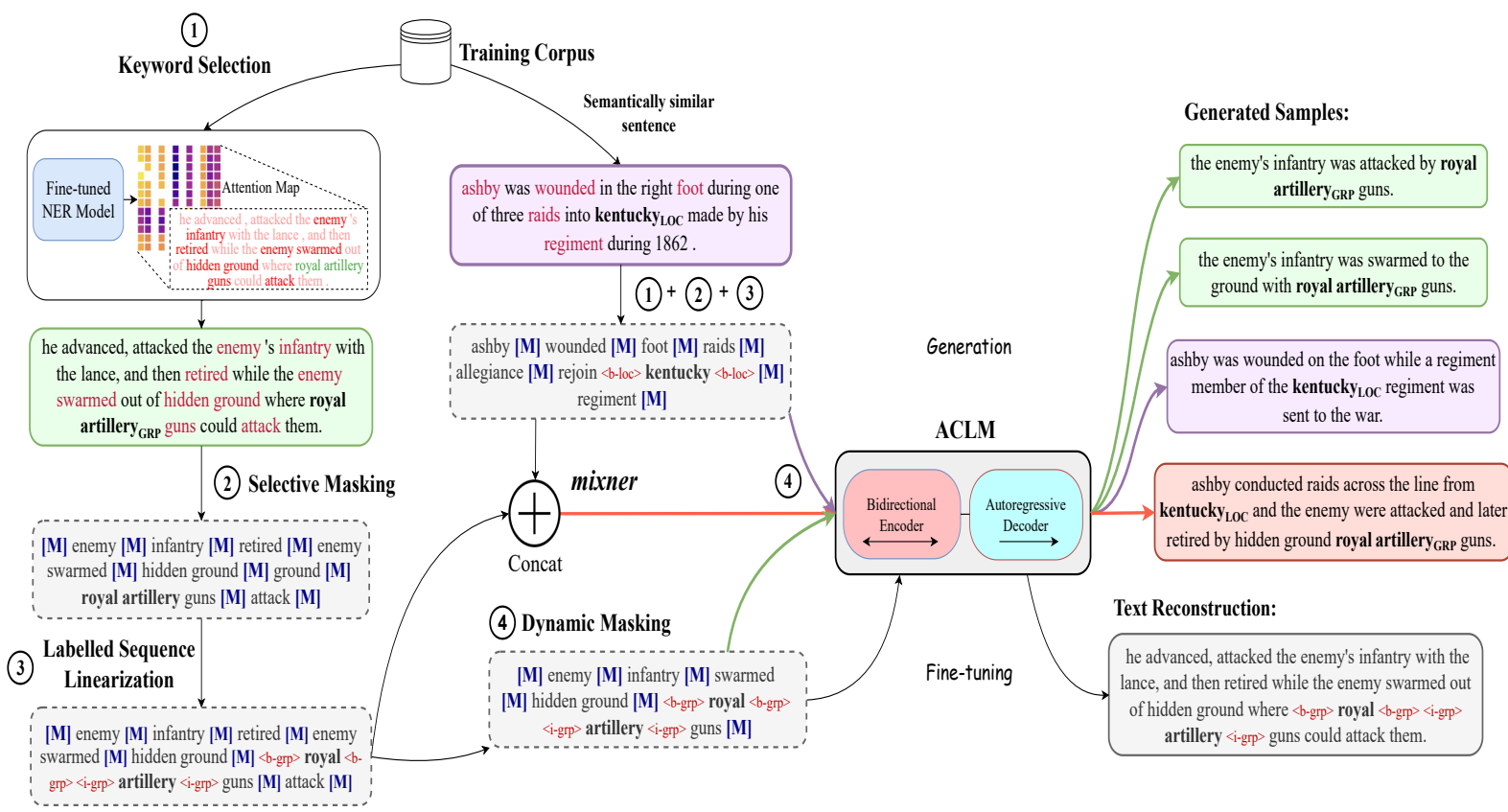
Sreyan Ghosh[1], Utkarsh Tyagi[1], Manan Suri[2], Sonal Kumar[1], S. Ramaneswaran[3], Dinesh Manocha[1]

University of Maryland, College Park, USA[1], NSUT Delhi, India[2], NVIDIA, Bangalore, India[3]

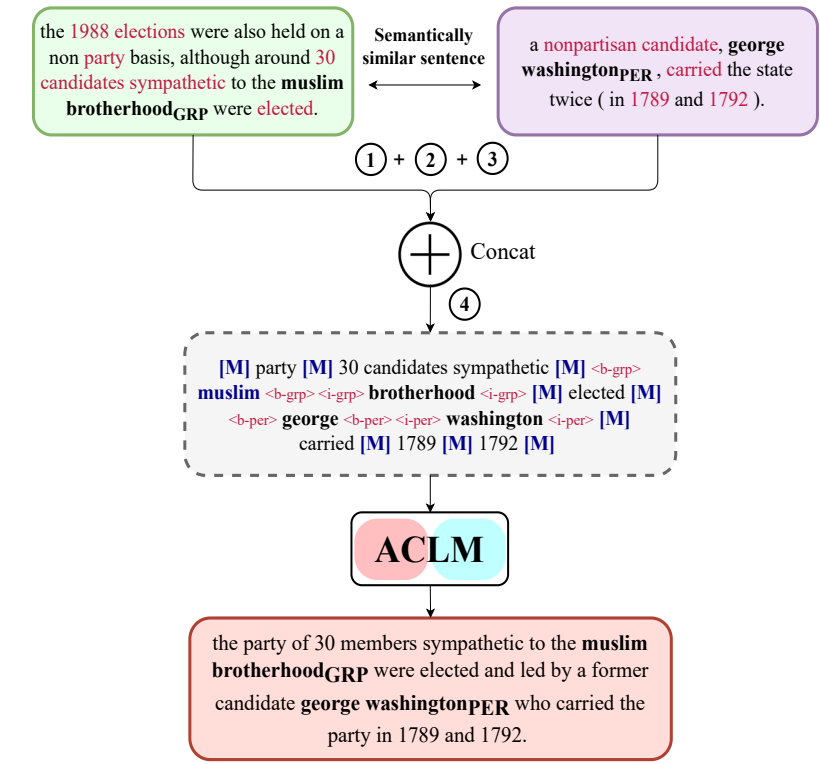## Motivation: Context-Entity Mismatch Issue in Data Augmentations

1. Complex Named Entity Recognition (NER) faces the issues of syntactically ambiguous and linguistically complex Named Entities (NEs), low context sentences, and emerging NEs which PLMs might not have seen in their pre-training data. These problems lead current NER data augmentation frameworks to generate incoherent augmentations with context-entity mismatch (where NEs do not match the surrounding context). They also impede a model from learning effective NE patterns, thereby leading to sub-optimal performance.

2. We propose **ACLM** (**A**ttention-guided **C**onditional **L**anguage **M**odel fine-tuning), a novel framework that optimizes BART on a novel text-reconstruction task based on *selective masking*. The resulting augmentations are coherent and diverse and outperform all our baselines.

Proposed Approach: Selective-Denoising with Attention Maps

Additionally, we propose *mixner* that mixes 2 semantically similar sentences before step 4 **to promote diversity.**

4 step sentence corruption for sequence-to-sequence denoising-based fine-tuning.

1. **Keyword Selection** using attention maps from fine-tuned transformer encoder **to preserve valuable hints**. The top-p% of tokens which are not NEs and have the highest attention score are selected.

2. **Selective Masking** by masking all other tokens except of the selected keywords in Step 1 and NEs.

3. **Labeled Sequence Linearization** by adding special tags arond each NE in a sentence **to provide explicit label supervision.**

4. **Dynamic Masking** by randomly masking some keywrods from Step 1 in each iteration **to promote length diversity.**

Original: *it was developed by a team led by former blizzard entertainment employees, some of whom had overseen the creation of the diablo series.*

LwTR: *it was developed by a makers led by, blizzard entertainment, some of whom had elevation the serving of the diablo 12th.*

MELM: *it was developed by a team led by former blizzago games employees, some of whom had overseen the creation of the hablo series.*

ACLM: *blizzard entertainment employees have overseen the production of the animated films, including the production of the diablo series.*

ACLM w/ mixner: *the team of the blizzard entertainment had overseen the creation of the game diablo and many of its workers founded pyro studios in the early 1960s.*

Results: ACLM outperforms all baselines Qualitatively and Quantitatively

#Gold	Method	MONOLINGUAL											CROSS-LINGUAL				
		En	Bn	Hi	De	Es	Ko	Nl	Ru	Tr	Zh	Avg	En→Hi	En→Bn	En→De	En→Zh	Avg
100	Gold-only	29.36	14.49	18.80	37.04	36.30	12.76	38.78	23.89	24.13	14.18	24.97	16.36	12.15	29.71	0.31	14.63
	LwTR	48.60	20.25	29.95	48.38	44.08	35.09	43.00	39.22	30.58	27.70	36.68	32.36	24.59	46.05	2.11	26.28
	DAGA	16.24	5.87	10.40	32.44	27.78	19.28	15.44	11.14	16.17	10.33	16.51	4.54	3.28	14.21	0.13	5.54
	MELM	40.12	6.22	27.84	43.94	37.45	34.10	37.82	32.38	20.13	25.11	30.51	26.37	20.33	34.32	2.71	20.93
	ACLM *only entity*	14.06	17.55	19.60	29.72	38.10	31.57	38.47	27.40	35.62	26.34	27.84	21.72	16.55	30.93	1.58	17.69
	ACLM *random*	43.59	20.13	28.04	45.83	42.27	33.64	41.82	38.20	36.79	25.99	35.63	29.68	21.64	45.27	3.05	24.91
	ACLM (ours)	48.76	23.09	33.53	48.80	44.14	38.35	46.22	39.48	37.20	35.12	39.47	32.52	23.91	46.48	3.58	26.62
200	Gold-only	51.83	19.31	33.68	49.62	45.16	42.51	47.83	31.55	26.76	32.34	38.06	36.90	27.44	48.70	3.76	29.20
	LwTR	52.88	23.85	34.27	50.31	47.01	42.77	52.01	40.18	35.92	30.57	40.98	40.07	32.36	48.95	6.04	31.85
	DAGA	33.30	17.12	19.58	35.10	33.56	26.50	38.04	29.83	23.35	25.66	28.20	18.92	14.37	29.32	1.79	16.10
	MELM	47.83	5.47	29.67	45.85	42.08	36.62	49.47	41.84	31.25	32.27	36.24	27.55	18.80	41.10	6.21	23.41
	ACLM *only entity*	50.06	25.58	37.78	50.95	48.21	43.39	48.46	34.87	34.92	28.20	40.24	30.76	22.53	44.17	6.50	25.99
	ACLM *random*	52.69	35.26	39.83	51.14	48.70	42.19	48.71	39.68	37.26	34.22	42.96	36.52	27.19	47.73	7.12	29.64
	ACLM (ours)	54.99	38.39	40.55	53.36	49.57	44.32	53.19	43.97	39.71	39.31	45.74	45.22	36.64	54.51	8.55	36.23
500	Gold-only	55.51	34.6	38.66	55.95	51.52	48.57	50.97	45.14	38.83	38.84	45.86	35.93	25.64	50.13	7.23	29.73
	LwTR	56.97	35.42	37.83	55.91	54.74	49.36	56.10	46.82	39.00	38.55	47.07	43.14	34.60	51.61	11.40	35.19
	DAGA	44.62	22.36	24.30	43.02	42.77	36.23	47.11	30.94	30.84	33.79	35.60	26.50	21.52	37.89	4.82	22.68
	MELM	52.57	9.46	31.57	53.57	46.40	45.01	51.90	46.73	38.26	39.64	41.51	34.97	27.17	44.31	7.31	28.44
	ACLM *only entity*	57.55	35.69	35.82	56.15	53.64	50.20	53.07	46.40	41.58	38.65	46.87	35.48	29.37	49.10	7.99	30.48
	ACLM *random*	57.92	38.24	39.33	57.14	53.24	49.81	55.06	48.27	42.22	40.55	48.18	41.72	32.16	52.27	13.63	34.95
	ACLM (ours)	58.31	40.26	41.48	59.35	55.69	51.56	56.31	49.40	43.57	41.23	49.72	44.36	35.59	54.04	16.27	37.57
1000	Gold-only	57.22	30.20	39.55	60.18	55.86	53.39	60.91	49.93	43.67	43.05	44.40	43.44	33.27	54.61	5.34	34.17
	LwTR	59.10	39.65	43.90	61.28	57.29	51.37	59.25	52.04	44.33	43.71	51.19	43.32	33.74	53.32	7.38	34.44
	DAGA	50.24	32.09	35.02	51.45	49.47	42.41	51.88	41.56	33.18	39.51	42.68	33.12	26.22	42.13	5.15	26.65
	MELM	53.48	6.88	37.02	58.69	52.43	50.50	56.25	48.99	36.83	38.88	44.00	35.23	25.64	46.50	8.22	28.90
	ACLM *only entity*	55.46	38.13	41.84	60.05	56.99	53.32	58.22	50.17	45.11	39.62	49.89	37.38	29.77	41.10	6.49	28.69
	ACLM *random*	58.87	41.00	46.27	61.19	57.29	53.61	59.52	52.77	45.01	43.60	51.91	43.96	34.14	53.37	7.25	34.68
	ACLM (ours)	60.14	42.42	48.20	63.80	58.33	55.55	61.22	54.31	48.23	45.19	53.74	44.59	35.70	56.74	8.94	36.49

Table 1: Monolingual and Cross-lingual Results: Improvement of 1% - 22%

#Gold	Method	En	Bn	Hi	De	Es	Ko	Nl	Ru	Tr	Zh	Avg
100 ×10	Gold-Only	56.21	35.66	42.16	55.71	54.98	45.14	57.48	46.13	44.40	30.72	46.86
	LwTR	55.65	38.47	43.44	54.71	53.95	44.78	56.50	46.93	45.41	31.56	47.14
	MulDA	46.87	29.25	34.52	45.92	45.55	33.91	48.21	38.65	35.56	27.33	38.58
	MELM	53.27	23.43	41.55	48.17	51.28	39.23	51.37	45.73	41.97	30.67	42.67
	ACLM (ours)	58.74	41.00	46.22	59.13	56.93	51.22	60.30	50.26	49.32	40.93	51.40
200 ×10	Gold-Only	58.67	39.84	46.34	59.65	58.50	50.70	59.21	51.66	47.12	40.98	51.42
	LwTR	51.78	35.93	38.87	52.73	51.59	42.55	54.49	43.99	41.23	35.19	44.83
	MulDA	48.89	31.45	36.76	48.41	48.30	39.78	51.09	42.01	35.98	31.65	41.43
	MELM	52.53	24.27	40.10	49.69	52.42	43.56	47.28	44.35	40.62	34.28	47.45
	ACLM (ours)	59.75	42.61	48.52	61.49	59.05	53.46	61.59	53.34	49.96	44.72	53.45
500 ×10	Gold-Only	61.10	40.94	48.20	61.67	59.84	54.56	62.36	53.33	48.77	45.82	53.66
	LwTR	59.09	38.37	43.80	59.37	57.76	50.38	60.42	51.00	46.53	42.97	50.96
	MulDA	51.79	30.67	35.79	51.87	50.92	43.08	53.95	44.61	38.86	36.72	43.83
	MELM	58.67	26.17	41.88	53.05	57.26	51.97	61.49	44.73	40.22	40.12	47.66
	ACLM (ours)	62.32	43.79	50.32	63.94	62.05	56.82	64.41	55.09	51.83	48.44	55.90
1000 ×10	Gold-Only	64.14	43.28	50.11	66.18	63.17	57.31	65.75	56.94	51.17	49.77	57.78
	LwTR	61.67	39.90	45.28	63.13	60.21	53.43	63.37	54.07	48.38	45.36	53.48
	MulDA	56.35	33.73	40.71	56.90	55.35	48.42	58.39	49.25	42.06	40.19	48.14
	MELM	51.55	30.27	42.61	61.45	59.31	53.00	48.48	44.35	48.93	44.71	52.24
	ACLM (ours)	64.50	46.59	52.14	67.65	64.02	59.09	67.03	57.82	53.25	50.60	58.27

Table 2: Multi-lingual Results: Improvement of 1% - 21%

#Gold	Method	CoNLL	BC2GM	NCBI	TDMSci	Avg
200	Gold-Only	79.11	50.01	72.92	47.20	62.31
	LwTR	82.33	52.78	72.15	51.65	64.73
	DAGA	76.23	47.67	71.14	48.03	60.77
	MELM	77.10	54.05	70.12	46.07	61.83
	ACLM (ours)	82.14	58.48	74.27	56.83	67.93
500	Gold-Only	84.82	55.56	75.75	47.04	65.79
	LwTR	85.08	60.46	78.97	60.74	71.31
	DAGA	81.82	51.23	78.09	57.66	67.20
	MELM	83.51	56.83	75.11	57.80	68.31
	ACLM (ours)	84.26	62.37	80.57	61.77	72.24

Table 3: Results on 4 other domains

#Gold	Method	Perplexity(↓)	Diversity-E(↑)	Diversity-N(↑)	Diversity-L(↑)
200	LwTR	137.01	30.72	16.46	0.0
	MELM	83.21	94.85	0.0	0.0
	ACLM (ours)	80.77	35.64	22.48	5.67
500	LwTR	129.349	30.07	16.22	0.0
	MELM	82.31	94.37	0.0	0.0
	ACLM (ours)	57.68	44.12	41.16	5.82
1000	LwTR	131.20	29.85	16.55	0.0
	MELM	82.64	95.13	0.0	0.0
	ACLM (ours)	62.00	50.10	34.84	5.40

Table 4: Qualitative Results

Code: https://github.com/Sreyan88/ACLM