BioAug

Conditional Generation based Data Augmentation for Low-Resource Biomedical NER

Sreyan Ghosh*, Utkarsh Tyagi*, Sonal Kumar*, Dinesh Manocha

University of Maryland, College Park

Introduction

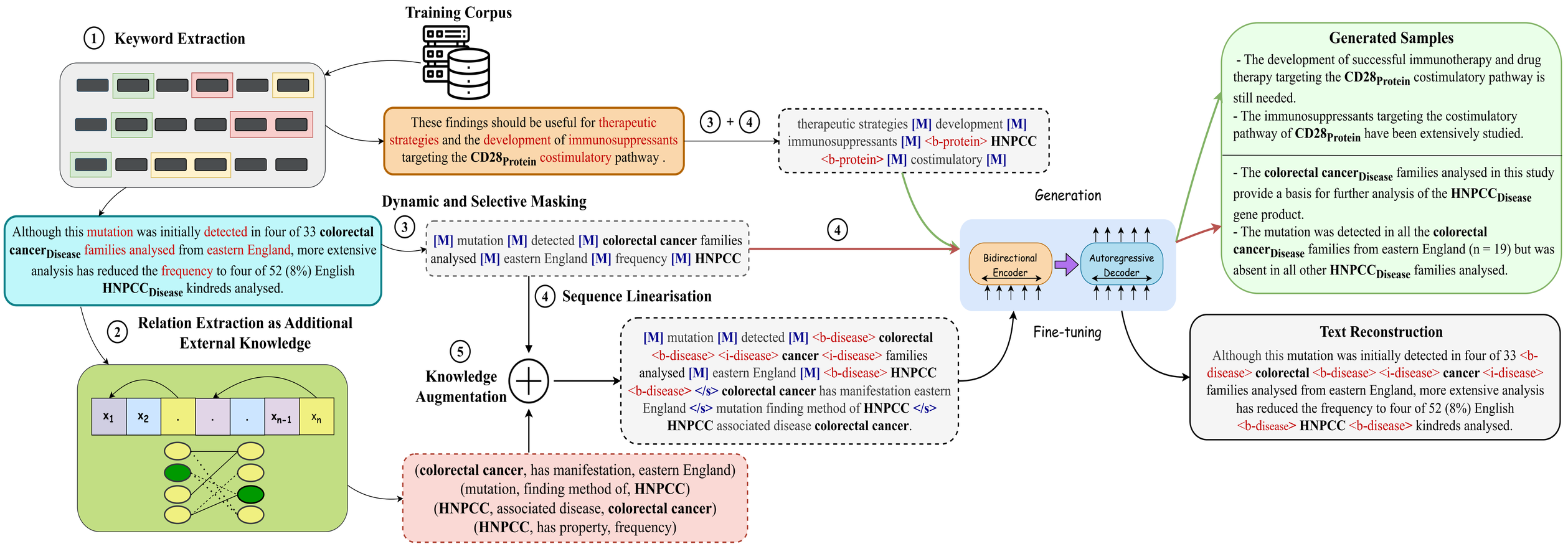
Problems in Biomedical Named Entity Recognition (**BioNER**):

1. Severe data scarcity
2. Lacks **high-quality** labelled data
3. Specialized and expert **knowledge** required for annotation
4. Lack of **factual** and **diverse** augmentations for BioNER

We present **BioAug**, a novel data augmentation framework for low-resource BioNER. BioAug, built on **BART**, is trained to solve a novel text reconstruction task based on **selective masking** and **knowledge augmentation**. Post training, we perform conditional generation and generate **diverse** augmentations conditioning BioAug on selectively corrupted text similar to the training stage.

BioAug – Sentence Corruption

1. **Keyword Extraction:** We first extract keywords from our sentence that provides contextually relevant knowledge about the target NE.
2. **Relation Extraction as Additional External Knowledge**: We add knowledge facts to the corrupted sentence during fine-tuning in the form of NE-NE, E-E, NE-E relation triplets.
3. **Dynamic & Selective Masking:** Remove the entities that overlap with the original NE and randomly select e% of the remaining entities. e is sampled from a Gaussian distribution.
4. **Sequence Linearization:** Add label tokens before and after each NE token and treat that as normal context in the sentence.
5. **Knowledge Augmentation:** The triples belonging to NE and entities left in the sentence are concatenated to the masked and linearized sentence.



Results

BioAug outperforms all baselines quantitatively and qualitatively

#Size	Model	BC2GM	BC5DR	NCBI	EBMNLP	JNLPBA	Avg.
100	Gold Only	56.94	74.90	72.99	18.81	44.37	53.60
	DAGA	38.63	60.96	58.26	17.48	43.85	43.84
	MulDA	39.67	62.35	59.56	20.32	45.66	45.51
	SR-UMLS	54.83	75.64	68.35	21.68	55.66	55.23
	MELM	48.56	74.70	65.74	24.64	50.32	52.79
	BioAug *(ours)*	**60.17**	**77.58**	**75.14**	**27.35**	**60.00**	**60.05**
200	Gold Only	62.16	76.08	76.02	23.96	54.26	58.50
	DAGA	48.95	68.69	70.92	23.53	53.58	53.13
	MulDA	50.11	69.35	72.28	25.37	55.28	54.48
	SR-UMLS	62.88	78.18	74.43	27.14	63.59	61.24
	MELM	58.78	79.06	73.49	21.19	58.18	58.14
	BioAug *(ours)*	**67.17**	**80.30**	**78.33**	**29.66**	**65.40**	**64.17**
500	Gold Only	65.97	82.55	80.18	31.48	62.04	64.44
	DAGA	53.95	76.60	78.70	32.41	61.72	60.68
	MulDA	54.92	78.04	79.92	33.53	62.63	61.81
	SR-UMLS	65.43	82.70	79.16	32.92	65.36	65.11
	MELM	58.78	81.19	75.49	32.26	61.64	61.87
	BioAug *(ours)*	**70.61**	**84.48**	**80.64**	**37.94**	**68.07**	**68.35**
All	Gold Only	82.33	89.01	87.33	42.98	74.36	75.20
	DAGA	79.62	86.69	85.15	42.46	72.52	73.29
	MulDA	80.21	87.55	86.93	44.54	73.78	74.60
	SR-UMLS	82.18	88.48	84.66	45.75	74.93	75.20
	MELM	81.46	89.18	83.95	40.38	73.82	73.76
	BioAug *(ours)*	**83.83**	**89.33**	**88.14**	**47.26**	**75.49**	**76.81**

#Gold	Method	Perplexity(↓)	Diversity(↑)	Diversity-L(↑)
100	SR-UMLS	115.76	14.65	2.38
	MELM	110.50	15.83	0.0
	BioAug *(ours)*	**39.69**	**47.88**	**9.074**
200	SR-UMLS	110.23	15.33	2.56
	MELM	97.78	18.65	0.0
	BioAug *(ours)*	**32.45**	**45.67**	**9.67**
500	SR-UMLS	102.55	14.98	2.42
	MELM	94.65	14.87	0.0
	BioAug *(ours)*	**31.14**	**44.72**	**10.17**

Generation Examples

Original	During an 18-month period of study 41 hemodialyzed patients receiving **[desferrioxamine]**$_{CHEMICAL}$ (10-40 mg/kg BW/3 times weekly) for the first time were monitored for detection of **[audiovisual toxicity]**$_{DISEASE}$.
SR-UMLS [16]	During an 18-month menstruation of survey 41 hemodialyzed patients find **[desferrioxamine]**$_{CHEMICAL}$ (10-forty magnesium/kilogram bw/triplet times weekly) for the showtime time were monitored for detecting of **[audiovisual perniciousness]**$_{DISEASE}$.
MELM [44]	During an 18-month period of study 41 hemodialyzed patients receiving **[glucoferriopamine]**$_{CHEMICAL}$ (10-40 mg/kg BW/3 times weekly) for the first time were monitored for detection of **[administration cycloone]**$_{DISEASE}$.
BioAUG *(ours)*	During a 12-month period , the study population consisted of 50 hemodialyzed patients receiving **[desferrioxamine]**$_{CHEMICAL}$, who were monitored for **[audiovisual toxicity]**$_{DISEASE}$.
Original	The authors assessed the **[safety]**$_{OUTCOME}$ and **[effectiveness]**$_{OUTCOME}$ of **[atomoxetine]**$_{INTERVENE}$ monotherapy compared with combined **[atomoxetine/fluoxetine therapy]**$_{INTERVENE}$ in a **[population of children and adolescents with ADHD and concurrent symptoms of anxiety]**$_{PATIENT}$.
SR-UMLS [16]	The generator tax the **[prophylactic]**$_{OUTCOME}$ and **[potency]**$_{OUTCOME}$ of **[atomoxetine]**$_{INTERVENE}$ monotherapy compared with combined **[atomoxetine/fluoxetine therapy]**$_{INTERVENE}$ in **[universe of minor and teen with ADHD and coincidental symptoms of anxiousness]**$_{PATIENT}$.
MELM [44]	The authors undertook this study to further evaluate the **[level of adenosine deaminase (ADA)]**$_{OUTCOME}$ in **[patients with chronic schizophrenia]**$_{PATIENT}$ treated with **[monotherapy of haloperidol, risperidone or clozapine]**$_{INTERVENE}$ and correlation between the **[ADA level]**$_{OUTCOME}$ with response to treatment.
BioAUG *(ours)*	The present study assessed the **[safety]**$_{OUTCOME}$ and **[effectiveness]**$_{OUTCOME}$ of **[atomoxetine]**$_{INTERVENE}$ therapy compared with a combination of **[atomoxetine and fluoxetine]**$_{INTERVENE}$ in the same **[population of children participants and adolescents with ADHD]**$_{PATIENT}$ (n = 60), and these two groups were also compared in terms of the frequency of these symptoms.

Paper

Code