ABEX: Data Augmentation for Low-Resource NLU via Expanding Abstract Descriptions

ACL 2024 Bangkok, Thailand

Sreyan Ghosh[1*], Utkarsh Tyagi[1*], Sonal Kumar[1], Chandra Kiran Everu[1], Ramaneswaran S, S Sakshi, Dinesh Manocha[1]
[1]University of Maryland, College Park, USA

UNIVERSITY OF MARYLAND

Introduction & Motivation

- Data augmentation has proven to be an effective approach for overcoming the data scarcity issue in low-resource NLU tasks with limited training samples.
- Generative data augmentation faces two major challenges: **diversity in generated augmentations** and **consistency with the underlying data distribution**.
- Current augmentation methods are either too conservative, by making small changes to the original text, or too aggressive, by creating entirely new samples. Additionally, they are prone to replicate biases and overfit specific linguistic patterns.

Main Contributions

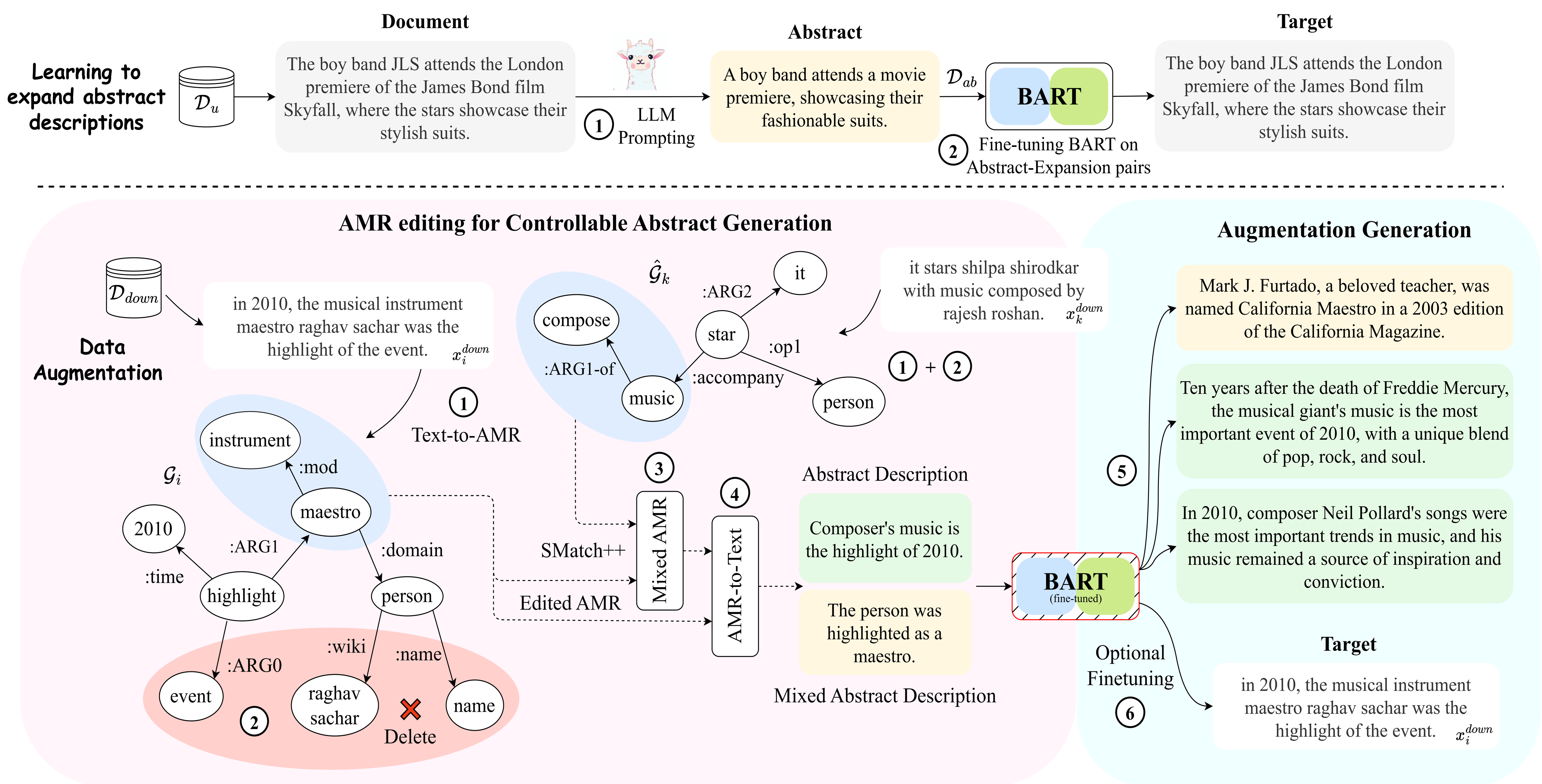
- We propose **ABEX** (**AB**stract and **EX**pand), a novel and effective generative data augmentation methodology for low-resource NLU. ABEX is based on a novel framework that first generates an abstract decsription of a document and then expands the abstract decsription.
- We propose a simple, controllable, and training-free method, based on editing AMR graphs, for generating abstract descriptions of documents from NLU datasets.
- We evaluate the efficacy of ABEX on 12 datasets across 4 NLU tasks under 4 low-resource settings and show that it outperforms most prior works quantitatively by 0.04% - 38.8%
- We contribute a large-scale synthetic dataset with ≈0.2 million abstract-expansion pairs to promote further research in this space.

Methodology

Abstarct Descriptions: We define an abstract description is a concise summary of a text, distilling it to its key concepts and themes while omitting non-essential details, effectively

Step 1: Learning to Expand Abstract Descriptions:
1) We synthesize a large-scale synthetic dataset D_{ab} with abstract-document pairs by prompting LLMs with unlabeled documents from D_{ab}.
2) We pre-train BART on this dataset with abstract as input and document as the target for learning to expand abstract descriptions.

Step 2: Synthetic Data Augmentation:
1) We convert the document into its AMR graph representation $\mathcal{G}_i$.
2) $\mathcal{G}_i$ then goes through multiple steps of deletion to obtain $\hat{\mathcal{G}}_i$.
3) We optionally retrieve a semantically similar document from Ddown, obtain its AMR graph $\mathcal{G}_k$, and replace subtrees in $\hat{\mathcal{G}}_i$ with similar subtrees in $\hat{\mathcal{G}}_i$.
4) $\hat{\mathcal{G}}_i$ is then converted back to text using an AMR-to-Text generator. The resultant text is now an abstract description of the document.
5) This abstract description is then passed to the fine-tuned BART for generating augmentations.
6) We optionally fine-tune the fine-tuned BART (from Step 1.) on abstract-document pairs from Ddown for downstream domain adaptation.



Quantitative Results

Model	Huffpost 100	Huffpost 200	Huffpost 500	Huffpost 1000	Yahoo 100	Yahoo 200	Yahoo 500	Yahoo 1000	IMDB 100	IMDB 200	IMDB 500	IMDB 1000	ATIS 100	ATIS 200	ATIS 500	ATIS 1000	MASSIVE 100	MASSIVE 200	MASSIVE 500	MASSIVE 1000
Gold	76.80	77.96	80.51	82.41	42.50	49.50	**55.47**	56.62	83.36	88.59	88.15	**89.47**	85.13	89.97	94.7	97.29	31.70	56.48	73.47	79.15
BackTrans	75.87	76.21	79.20	80.20	44.85	50.86	54.19	55.77	84.38	86.12	86.72	87.53	89.86	92.34	94.36	97.07	53.56	64.52	73.13	78.48
EDA	75.49	77.64	79.14	80.71	47.13	50.15	53.39	56.04	75.3	88.07	88.39	88.92	90.20	92.11	94.93	96.62	47.00	64.15	73.53	78.24
AEDA	77.65	76.88	80.31	81.10	45.61	51.52	54.22	56.02	82.30	88.25	86.95	89.33	89.07	91.89	96.73	97.63	51.04	66.81	75.15	79.11
AMR-DA	77.49	76.32	77.93	79.64	48.80	52.37	54.68	55.01	84.26	88.04	88.92	89.20	93.69	94.03	96.28	96.39	52.82	64.02	72.09	76.96
SSMBA	76.64	77.40	79.85	81.11	46.95	50.53	53.97	54.68	82.09	86.57	87.94	88.8	90.31	89.75	93.69	95.94	47.07	60.99	70.24	77.16
GENIUS	77.52	77.71	78.35	80.07	51.9	51.69	51.46	54.15	78.58	82.50	84.90	86.18	93.58	94.14	96.73	97.18	51.76	65.34	73.17	77.04
PromDA	77.83	77.90	77.65	81.06	52.61	52.13	53.40	56.27	84.21	88.24	88.30	88.65	-	-	-	-	-	-	-	-
PromptMix	-	-	-	-	-	-	-	-	-	-	-	-	92.68	94.25	94.81	96.95	52.60	64.53	74.26	76.87
ZeroGen	73.84	75.66	76.30	76.49	41.47	49.21	54.55	55.04	76.99	80.61	82.31	83.10	81.24	83.95	85.63	90.88	28.20	47.02	67.80	70.94
LLaMA-2_{13B}	73.59	75.19	76.82	77.94	40.37	46.25	52.14	53.62	80.72	83.59	85.62	85.81	82.80	81.72	89.11	91.05	30.88	49.19	70.52	71.80
GPT3Mix	57.87	61.80	66.12	69.46	31.60	32.98	50.33	52.93	81.04	84.14	86.27	87.69	76.91	81.75	85.36	85.36	25.91	46.72	68.99	72.57
ABEX-Abs	73.62	74.58	76.27	78.42	35.87	37.93	48.47	50.36	74.69	80.28	82.66	82.51	78.53	80.27	83.54	86.49	30.71	51.62	68.88	75.26
ABEX-stage-2	74.61	77.26	78.17	80.28	49.81	50.02	51.62	53.74	82.69	85.36	87.22	87.45	90.71	92.36	96.75	96.68	50.47	65.38	73.29	76.25
ABEX-stage-1	77.45	79.24	81.63	83.58	52.46	53.26	54.77	**57.13**	84.35	88.16	88.30	89.17	91.66	94.83	96.79	96.45	52.51	65.63	73.94	79.41
ABEX *(ours)*	**78.66**	**79.30**	**81.82**	**84.03**	**53.20**	**53.52**	54.81	57.11	**85.18**	**88.72**	**89.05**	89.28	**94.28**	**95.71**	**97.33**	**97.92**	**55.03**	**66.85**	**75.44**	**80.36**
	±0.72	±0.05	±0.13	±0.42	±0.56	±0.24	±0.51	±0.01	±0.73	±0.12	±0.10	±0.12	±0.54	±0.78	±0.45	±0.24	±1.34	±0.02	±0.24	±0.85

Result comparison on Sequence Classification. ABEX outperforms prior methods by 0.04% - 29.12%.

Model	CoNLL-2003 100	CoNLL-2003 200	CoNLL-2003 500	CoNLL-2003 1000	MultiCoNER 100	MultiCoNER 200	MultiCoNER 500	MultiCoNER 1000	OntoNotes 100	OntoNotes 200	OntoNotes 500	OntoNotes 1000
Gold-only	52.89	66.53	70.43	80.15	15.86	24.91	52.69	57.03	16.37	27.7	61.46	61.82
LwTR	65.48	73.24	81.45	83.74	42.23	50.22	51.0	54.67	46.18	**51.47**	54.87	62.67
DAGA	53.91	51.63	54.68	82.05	19.11	36.71	31.39	42.13	33.29	43.07	54.64	61.15
MELM	56.89	62.23	79.05	81.90	16.62	30.96	46.27	49.01	11.94	31.55	45.68	54.97
GENIUS	67.85	58.2	80.36	76.87	42.33	47.77	55.70	51.06	45.44	48.69	52.27	56.59
PromDA	66.30	70.95	76.38	82.14	41.40	48.93	55.02	53.55	46.34	50.83	54.81	57.64
LLaMA-2_{13B}	53.39	68.71	73.95	79.22	39.82	45.36	50.60	55.68	40.61	43.29	53.72	57.88
GPT-NER	54.61	68.25	78.17	80.60	40.81	46.37	52.19	55.92	42.37	44.82	55.20	58.62
ABEX-Abs	54.18	65.52	72.36	79.40	24.62	35.28	44.71	47.90	30.76	35.26	43.28	50.60
ABEX-stage-2	68.22	71.15	77.02	82.41	41.25	48.73	54.14	54.36	45.85	47.92	55.88	57.62
ABEX-stage-1	68.74	72.09	78.51	83.22	41.28	49.44	54.73	55.60	46.82	45.71	56.63	59.25
ABEX *(ours)*	**70.16**	**73.67**	**83.58**	**84.20**	**43.05**	**51.75**	**56.03**	**58.41**	**48.76**	51.38	**61.85**	**63.14**
	±0.86	±0.37	±1.27	±0.31	±0.67	±1.32	±0.24	±1.24	±1.23	±0.06	±0.26	±0.35

Result comparison on NER. ABEX outperforms prior methods by 0.33% - 36.82%.

Results for Question-Answering and Sentence Similarity on paper!

Qualitative Results

Original	Both countries are full members of the european union$_{GRP}$ and nato$_{GRP}$.	Coherence	Context Diversity	Label Consistency
LwTR	1. Both countries are full 1856 of the european union and colorado. 2. Both countries are full members of the european historic and nato inspired	✗	✗	✗
DAGA	1. It is by as european is three rubber singer age her. 2. he also of a and european to.	✗	✗	✗
MELM	1. Both countries are full members of the national socialist and international. 2. Both countries are full members of the national states and international.	✓	✗	✓
GENIUS	1. The european union, nato, paul getty museum and metropolitan museum of art 2. The european union is in trouble. nato, the EU is in the crosshairs.	✗	✗	✓
PromDA	1. European union and nato resisted invasions of their countries. 2. They are also members of the european nato and european union &&.	✗	✓	✓
Abstract Description	Countries in unions and military members.	-	-	-
ABEX-ft	1. 23 countries are a part of the european union and nato military. 2. The groups and militaries were the only equivalents of nato, one of whom were under the political influence	✓	✓	✓
ABEX-pt	1. European Union$_{GRP}$ member states such as Romania$_{LOC}$, Serbia$_{LOC}$, and Spain$_{LOC}$ are part of the european union$_{GRP}$ with the nato$_{GRP}$ military, but may not join the union due to a lack of agreement and their ability to compete in the union. 2. Nine member states of the european union$_{GRP}$ are part of the country's nato$_{GRP}$ military force, with participation restricted to minor groups and no restrictions on freedom of movements, and have historically been part of North Dakota$_{LOC}$'s nanton system.	✓	✓	✓
ABEX (*ours*)	1. The Netherlands$_{LOC}$ is a member of the european union$_{GRP}$, joined in 1969; the Netherlands$_{LOC}$ is also a member of nato$_{GRP}$ with an observer status. 2. The european union$_{GRP}$ is composed of 12 countries, with the majority of them being members of the nato$_{GRP}$, and the union's member states.	✓	✓	✓

Comparison of augmentations on the MultiCoNER dataset.

Method	P(↓)	D(↑)	D-L(↑)	P(↓)	D(↑)	D-L(↑)
	100			500		
EDA	135.12	103.49	10.63	147.06	120.69	12.07
SSMBA	86.13	126.66	17.58	103.92	134.44	19.12
AEDA	105.92	49.72	6.55	106.87	50.56	6.99
BackTrans	77.17	34.02	19.39	74.98	47.22	20.91
GPT3-Mix	90.50	124.02	23.55	85.49	134.08	26.98
GENIUS	32.88	156.50	27.95	32.71	159.49	28.13
AMR-DA	68.22	68.73	2.58	64.95	75.15	2.92
LWTR	152.69	101.95	11.39	137.03	109.02	11.64
DAGA	66.46	54.59	14.91	120.74	69.32	10.74
MELM	69.13	113.39	12.91	83.43	116.59	11.30
ABEX-stage-1 *(ours)*	27.46	**190.87**	27.74	26.48	**217.29**	17.88
ABEX *(ours)*	**28.05**	124.91	**29.73**	**27.09**	130.25	**31.37**

Comparison of perplexity (P), token diversity (D), and length diversity (D-L).

Code and Data