DALE : Generative Data Augmentation for Low Resource Legal NLP

Sreyan Ghosh[1], Chandra Kiran Everu[1], Sonal Kumar[1], S Rameshwaram[1], S Sakshi[2], Utkarsh Tyagi1[1], Dinesh Manocha[1]
[1]University of Maryland, College Park, [2]UMass Amherst

EMNLP 2023

Introduction

Motivation: Legal documents, with its complex semantics, morphology, and syntax, does not benefit from data augmentations that merely rephrase the source sentence.

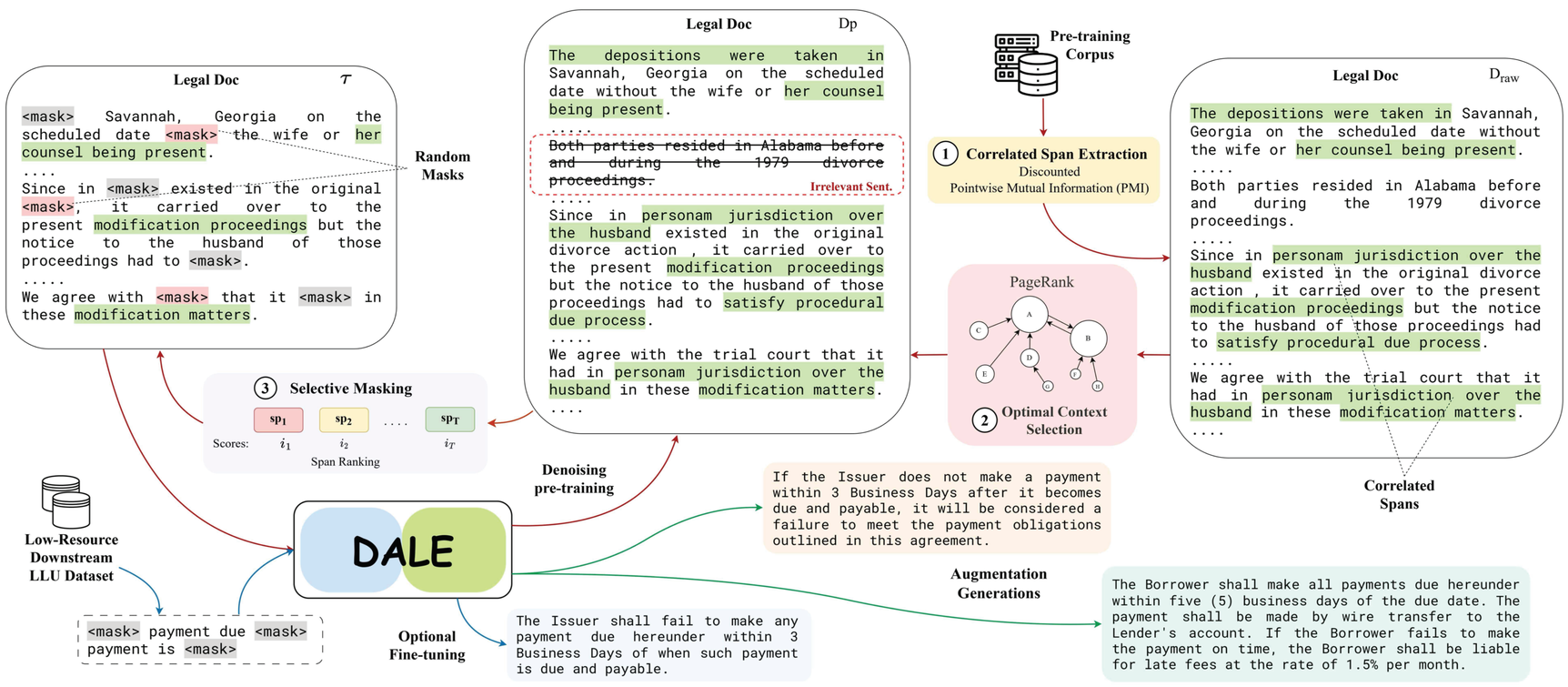
Solution: We present **DALE**, a novel and effective generative **D**ata **A**ugmentation framework for lowresource **LE**gal NLP. DALE, built on an EncoderDecoder Language Model, is pre-trained on a novel unsupervised text denoising objective based on selective masking - our masking strategy exploits the domain-specific language characteristics of templatized legal documents to mask collocated spans of text.

		Preserves Hints	Avoids Randomness
Orig:	Did the superior court abuse its discretion in dismissing Morgans appeal for failure to exhaust administrative remedies ?		
RM:	<mask> abuse <mask> discretion <mask> Morgans appeal <mask> to exhaust administrative <mask>	✗	✗
GM:	<mask> abuse its discretion <mask> dismissing Morgans appeal <mask> to exhaust administrative <mask>	✗	✗
PMI:	Did the <mask> abuse its discretion in dismissing <mask> appeal for failure to exhaust <mask> ?	✗	✗
DM:	<mask> in dismissing Morgans <mask> to exhaust administrative <mask>?	✓	✓
	<mask> in failing to allow Hertz to intervene as a pro se plaintiff ? } Other sentences with the same co-occurring span	✓	✓
	<mask> in awarding attorneys fees to moore in the <mask> 12,560.37? }	✓	✓

Method

To modify the existing or introduce a novel context in legal documents while maintaining the formal legal style and plausibility of events in the generated context, DALE, like a legal practitioner, should possess both broad legal knowledge and knowledge of legalese.

1. **Correlated Span Extraction:** We extract all correlated spans from a legal corpus using a novel discounted PMI formulation.
2. **Optimal Context Selection:** We shorten a legal document by selecting only the top-k sentences that are the most relevant to the document and removing the rest.
3. **Ranking and Template Creation:** We rank all the spans based on their importance and length using our novel scoring metric. Finally, we create a template by retaining the top-p spans and masking all other spans with with added randomness.



Quantitative Results

#Gold	100	200	500	1000	100	200	500	1000	100	200	500	1000	100	200	500	1000	100	200	500	1000
Dataset	**OTS-TOPICS**				**EUR-LEX**				**ECtHR-A**				**ECtHR-B**				**UNFAIR-ToS**			
Gold-only	0.10	11.47	51.16	53.87	8.68	4.30	10.32	42.26	25.26	27.30	17.14	31.52	37.69	47.47	44.89	50.98	0.10	33.88	70.02	76.21
EDA	9.72	38.43	37.56	46.99	12.11	22.93	49.26	51.54	10.10	35.64	41.91	49.67	43.01	48.70	56.32	59.40	13.93	26.31	72.15	78.14
Legal-EDA	10.10	39.15	40.40	50.48	12.45	23.61	51.24	53.27	12.24	36.75	43.89	52.93	43.86	54.72	57.71	61.53	15.86	27.54	72.98	78.69
SSMBA	10.41	15.28	47.31	52.63	4.10	21.32	45.67	48.70	7.55	18.10	34.39	37.58	35.32	45.43	48.08	52.65	6.53	18.21	63.96	68.59
AEDA	14.06	52.63	60.29	72.32	3.07	33.33	50.33	52.21	28.12	30.94	32.29	45.48	39.15	50.85	50.48	51.26	8.08	52.34	70.48	73.67
SMERTI	3.41	17.90	57.26	60.54	6.62	27.86	44.45	47.68	28.51	22.61	23.43	38.59	38.43	51.02	52.07	53.71	20.46	47.31	59.38	69.27
BackTrans	8.26	37.44	47.47	50.85	5.03	19.63	37.86	42.65	14.73	17.37	35.36	39.41	37.61	49.88	50.77	52.83	12.84	39.28	46.51	62.64
C-MLM	3.85	17.95	58.54	61.45	7.17	28.21	45.04	47.85	27.95	23.24	23.89	39.23	39.46	52.17	53.26	54.68	20.42	48.52	59.87	69.62
GENIUS	25.58	54.31	63.71	67.29	5.79	34.03	53.19	57.95	28.68	28.66	36.38	43.67	40.40	44.03	50.54	54.29	11.20	47.18	67.71	75.79
ChatGPT	23.42	53.31	62.17	65.87	5.52	33.22	52.21	56.45	27.52	27.89	34.03	41.83	39.61	43.12	49.76	53.87	10.78	44.62	65.87	72.91
Falcon	12.36	37.84	48.66	51.74	5.11	22.02	46.19	49.03	17.68	20.39	35.81	38.62	36.12	46.53	47.27	53.85	5.44	16.10	62.82	67.51
DALE-BART	25.77	54.01	58.29	68.04	12.32	34.39	53.65	56.27	23.01	35.68	40.13	52.47	43.91	52.76	54.58	60.24	18.43	46.60	68.21	75.04
DALE-pt	24.58	52.17	58.18	69.97	11.50	29.51	51.63	53.12	24.19	33.87	40.87	48.85	42.97	51.67	51.63	59.23	18.54	47.59	63.21	73.56
DALE-ft	24.63	53.22	59.64	70.15	11.61	33.54	52.38	57.62	24.21	34.76	41.78	51.65	43.33	53.74	55.12	60.95	19.11	48.71	67.42	74.86
DALE *(ours)*	**33.91**	**61.23**	**71.56**	**73.24**	**13.50**	**37.93**	**55.99**	**59.45**	**29.43**	**37.57**	**44.38**	**55.72**	**46.72**	**56.13**	**59.18**	**64.01**	**22.32**	**54.62**	**74.84**	**82.98**

#Gold	100	200	500	1000	100	200	100	200
Dataset	**CaseHOLD**				**BUILD-RR**		**ContractNLI**	
Gold-only	33.92	66.38	70.06	70.80	74.62	78.24	72.03	82.06
EDA	56.38	64.71	66.42	69.45	77.33	81.83	73.92	75.40
AEDA	57.96	65.10	69.12	70.05	77.95	82.01	77.24	83.02
SSMBA	62.01	67.65	69.59	69.75	77.77	81.66	76.27	82.93
SMERTI	56.52	64.13	69.15	69.85	77.42	80.65	76.23	81.95
BackTrans	55.69	65.72	69.29	69.74	77.59	81.08	75.98	81.19
GENIUS	55.84	61.37	64.17	68.20	78.99	79.30	77.28	81.28
ChatGPT	54.67	60.83	62.57	67.59	77.32	78.37	76.29	80.10
Falcon	52.57	58.76	62.41	63.22	75.11	77.61	75.17	77.54
DALE-BART	61.21	66.09	67.91	70.64	78.59	80.01	76.56	81.27
DALE-pt	59.25	65.69	67.81	69.70	78.15	79.01	76.97	80.55
DALE-ft	60.31	66.56	68.46	70.15	78.50	79.72	77.10	81.73
DALE *(ours)*	**63.71**	**68.14**	**71.53**	**72.70**	**81.83**	**83.04**	**79.26**	**85.13**

#Gold	100	200	500	1000	100	200	500	1000
Baselines	**EDGAR**				**INDIAN LEGAL NER**			
Gold-only	0.75	0.27	34.86	57.84	8.41	13.61	33.28	42.6
LwTR	22.10	36.84	50.33	54.15	12.53	17.87	35.54	44.15
DAGA	13.21	24.54	36.15	42.58	5.13	14.52	26.13	31.74
MulDA	8.17	21.33	42.61	50.16	13.75	19.28	31.96	40.69
MR	19.13	36.62	50.95	58.33	18.62	25.26	43.14	49.68
MELM	12.32	24.35	48.72	60.59	14.55	21.69	38.73	48.64
GENIUS	13.79	28.44	50.93	62.69	19.05	29.28	48.72	53.61
PromDA	10.10	27.31	45.77	55.62	16.46	26.91	45.34	44.62
ChatGPT	12.65	26.32	49.25	60.67	18.24	27.58	46.44	51.41
Falcom	11.24	25.71	48.69	59.84	18.11	26.23	43.05	49.38
DALE-BART	17.76	34.20	48.71	57.99	16.43	29.19	46.03	49.96
DALE-pt	18.38	33.12	47.67	53.67	17.25	27.86	45.57	48.28
DALE-ft	19.10	35.39	48.20	58.74	17.65	28.32	46.71	49.98
DALE *(ours)*	**23.65**	**39.82**	**55.99**	**64.32**	**21.31**	**32.47**	**49.93**	**54.27**

DALE outperforms all the baselines quantitatively on 13 datasets spanning 6 tasks and 4 low-resource settings.

EMNLP 2023
Paper: https://arxiv.org/pdf/2310.15799v1.pdf
Code: https://github.com/Sreyan88/DALE

Qualitative Results

Method	**Perplexity(↓)**	**Diversity(↑)**	**Diversity-L(↑)**	**Perplexity(↓)**	**Diversity(↑)**	**Diversity-L(↑)**
	200			500		
EDA	82.22	12.49	83.48	86.14	12.72	86.28
Legal-EDA	55.38	25.71	13.51	58.92	26.70	14.26
SSMBA	37.96	54.74	17.74	37.84	56.85	19.29
AEDA	26.93	2.17	176.68	27.05	13.67	145.13
SMERTI	28.56	56.84	13.76	29.20	59.62	14.58
BackTrans	27.94	45.05	27.62	27.85	49.05	28.62
C-MLM	50.39	41.04	23.85	51.69	44.86	25.69
GENIUS	24.37	106.08	226.65	24.65	105.04	278.64
GPT3-Mix	52.76	42.21	29.74	53.21	45.73	33.68
PromDA	174.67	65.69	15.74	187.68	73.93	16.84
LWTR	481.34	86.91	49.87	413.66	76.37	21.42
MR	82.72	75.65	29.23	79.65	81.46	32.76
MELM	211.94	12.49	83.48	183.23	12.72	86.28
ChatGPT	26.29	64.31	32.85	26.17	66.94	35.85
Falcon	45.24	13.64	17.63	44.97	15.74	18.59
DALE-BART	20.36	172.54	222.37	21.65	193.32	231.86
DALE-pt	58.09	66.99	**260.00**	60.12	59.84	**294.05**
DALE-ft	18.75	149.77	219.22	20.21	156.54	200.99
DALE *(ours)*	**18.63**	**175.38**	227.39	**18.44**	**194.20**	234.86

	UNFAIR ToS
Original	The most recent version of this agreement will be posted on the services under settings and also on gotinder.com, and you should regularly check for the most recent version.
EDA	recent version of this agreement will be posted on the services under settings and also on gotinder com and you should regularly check for the most recent version ✗ ✗ ✓
AEDA	the most ; recent version of ; this agreement will be posted on the , services under settings and also on gotinder.com . , and you should regularly check for the most recent version . , ✗ ✗ ✓
SMERTI	This most recent version of Windows will be posted on power under settings available on gotinder. , and you should regularly check our most recent version. ✗ ✗ ✗
GENIUS	The terms of this agreement will be contingent on the services they provide. For more information, please visit www.sos.gov. ✓ ✗ ✗
ChatGPT	The latest edition of this agreement will be made available on the services, specifically under the settings section and on gotinder.com. It is advisable to frequently review the most recent version. ✓ ✗ ✓
Falcon	The most recent version of this agreement will be posted on the services under settings and also on gotinder.com, and you should regularly check for the most recent version. ✓ ✗ ✓
DALE-pt	The most recent version of this agreement shall be accepted as the most recent amendment . ✓ ✗ ✗
DALE-ft	the most recent version of this agreement will be posted on the services under settings and also on gotinder.com, and you should regularly check for the most most recent versions. ✓ ✗ ✓
DALE	The most recent version of this agreement will be posted on the services's website at https://www.adr.nianticlabs.com/ where you can download and view the services, and you should be aware that this is not a guarantee that the services will be up to code or up to date, and we reserve the right to discontinue using the services at any time. ✓ ✓ ✓

DALE Generations. Pink signifies the change