SDA: Simple Discrete Augmentation for Contrastive Sentence Representation Learning

Dongsheng Zhu[1,3,*], Zhenyu Mao[2,*], Jinghui Lu[2,*], Rui Zhao[2], Fei Tan[2]

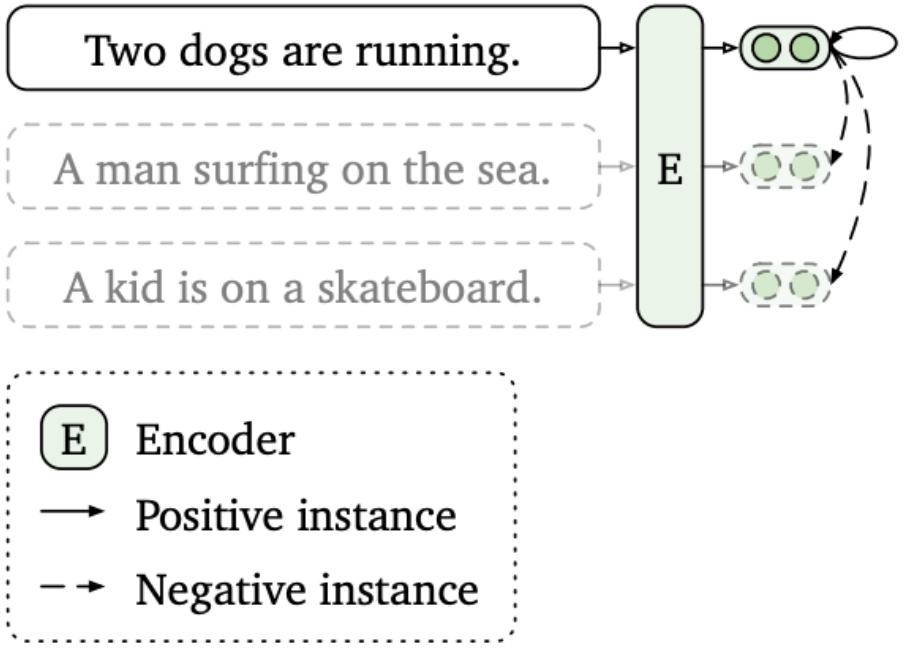
[1] Baidu Inc. [2] SenseTime Research [3] Fudan University
* Equal contribution

Speaker: Dongsheng Zhu

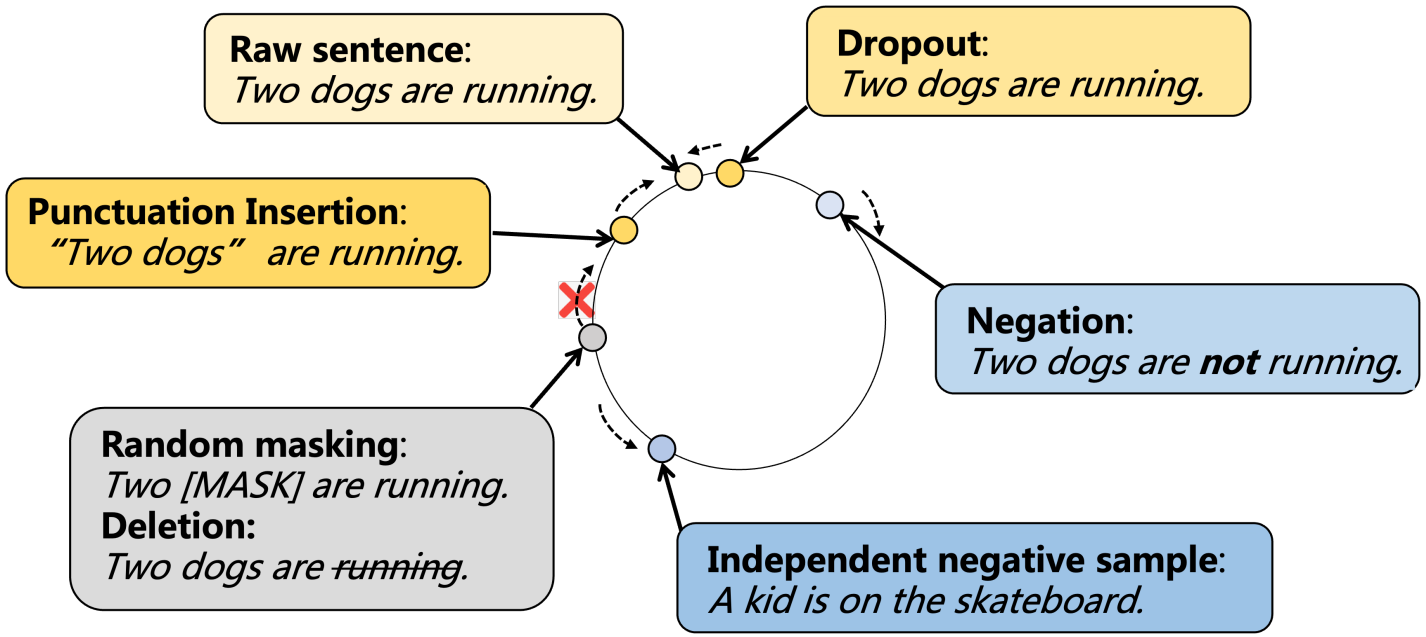
Conference: LREC-COLING 2024

Unsupervised SimCSE

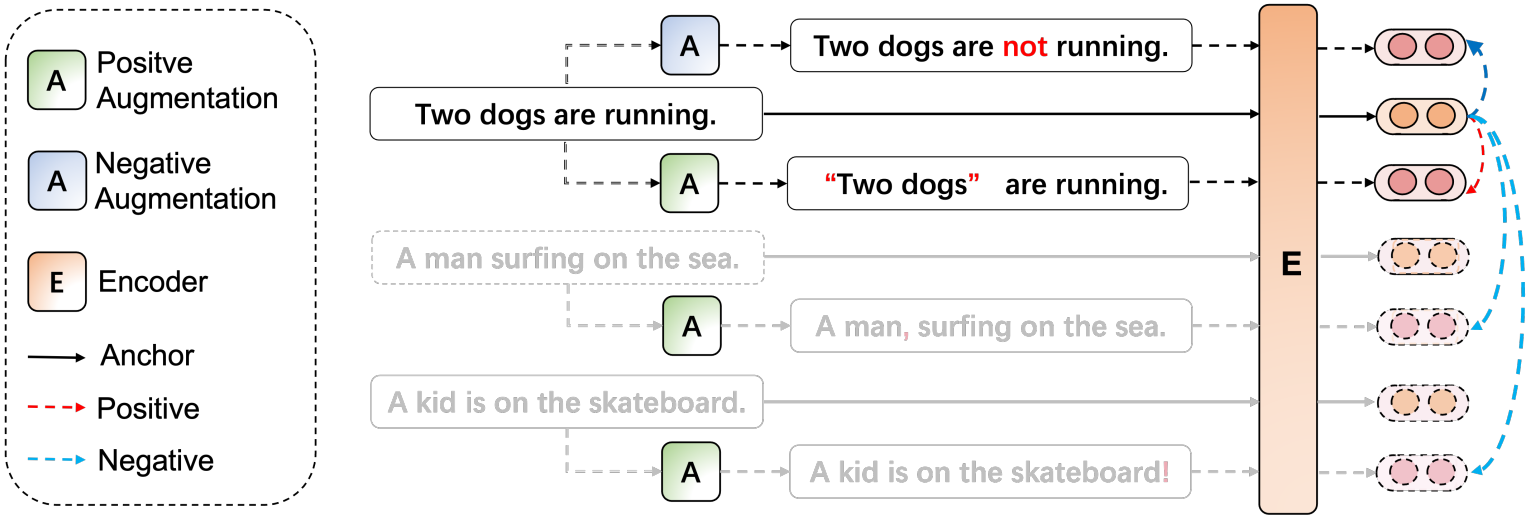
- Different hidden dropout masks in two forward passes.
- Positive pairs: embeddings of the same sentence with different dropout masks.
- Negative pairs: embeddings of other sentences from the same batch.



Visualization of Different Augmentation Methods



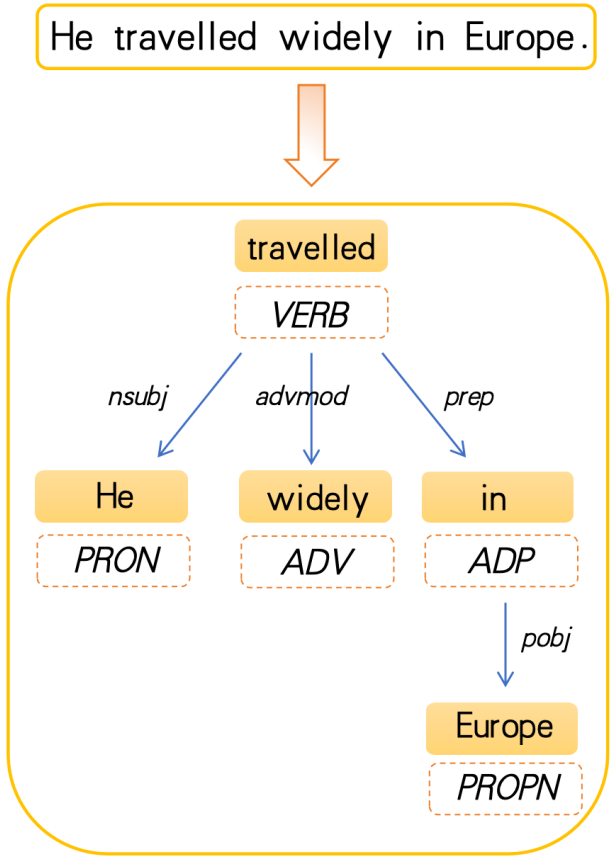
Overview and Example



- Punctuation Insertion (PI)
- Modal Verbs (MV)
- Double Negation (DN)

Method	Sentence
None	He travelled widely in Europe.
PI	He , travelled widely in Europe.
MV	He **must have** travelled widely in Europe.
DN	**It is not the fact that** he **didn't** travel widely in Europe.
Negation	he **didn't** travel widely in Europe.

Dependency Parsing and its Representation



Hyper-parameter δ with Negation

Raw sentence Positive

$$l_i = -\frac{e^{\mathrm{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} e^{\mathrm{sim}(\mathbf{h}_i, \mathbf{h}_j^+)/\tau}}$$

Negative

$$l_i = -\frac{e^{\mathrm{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} e^{\mathrm{sim}(\mathbf{h}_i, \mathbf{h}_j^+)/\tau} + e^{[\mathrm{sim}(\mathbf{h}_i, \mathbf{h}_i^-) - \delta]/\tau}}$$

Negation

Main Results

Methods	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
				BERT Models				
IS-BERT$_{base}$ ♢	56.77	69.24	61.21	75.23	70.16	69.21	64.25	66.58
CT-BERT$_{base}$ ♢	61.63	76.80	68.47	77.50	76.48	74.31	69.19	72.05
ConSERT$_{base}$ ♠	64.64	78.49	69.07	79.72	75.95	73.97	67.31	72.74
SimCSE-BERT$_{base}$ ♢	68.40	82.41	74.38	80.91	78.56	76.85	72.23	76.25
ESimCSE-BERT$_{base}$ ♣	69.79	83.43	75.65	82.44	79.43	79.44	71.86	77.43
ArcCSE-BERT$_{base}$ ♠	72.08	84.27	76.25	82.32	79.54	79.92	72.39	78.11
DiffCSE-BERT$_{base}$ ♡	**72.28**	84.43	76.47	83.90	80.54	**80.59**	71.23	78.49
Our Methods								
Punctuation Insertion	71.92	84.38	**76.84**	**83.92**	80.45	80.25	74.26	**78.86**
Modal Verbs	71.35	84.45	76.60	83.77	**80.57**	80.31	**74.85**	78.84
Double Negation	71.23	**84.49**	75.88	83.34	79.37	79.67	74.32	78.33
Ensemble	72.31±0.38	83.66±0.64	76.59±0.38	84.10±0.20	80.41±0.42	80.17±0.52	72.78±0.49	78.57±0.43
				RoBERTa Models				
DeCLUTR-RoBERTa$_{base}$ ♢	52.41	75.19	65.52	77.12	78.63	72.41	68.62	69.99
SimCSE-RoBERTa$_{base}$ ♢	70.16	81.77	73.24	81.36	80.65	80.22	68.56	76.57
ESimCSE-RoBERTa$_{base}$ ♣	69.90	82.50	74.68	83.19	80.30	80.99	70.54	77.44
DiffCSE-RoBERTa$_{base}$ ♡	70.05	83.43	75.49	82.81	82.12	82.38	71.19	78.21
Our Methods								
Punctuation Insertion	70.92	83.59	76.87	**83.73**	**82.42**	**83.02**	74.89	79.35
Modal Verbs	**72.37**	**83.80**	77.51	83.58	82.29	82.98	74.69	**79.60**
Double Negation	71.07	83.56	**77.60**	83.38	81.59	81.82	**75.44**	79.21
Ensemble	72.64±0.62	83.45±0.38	76.90±0.17	83.56±0.21	81.82±0.38	82.76±0.31	74.77±0.6	79.37±0.07

Transfer Tasks

Method	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	Avg.
SimCSE-RoBERTa$_{base}$♢	81.18	86.46	94.45	88.88	85.50	89.80	74.43	85.81
ArcCSE-BERT$_{base}$♠	79.91	85.25	**99.58**	**89.21**	84.90	89.20	74.78	86.12
DiffCSE-RoBERTa$_{base}$♡	82.82	88.61	94.32	87.71	**88.63**	90.40	76.81	87.04
Our Methods								
Punctuation Insertion	**83.59**	87.79	93.81	88.10	87.81	**91.60**	**76.93**	**87.09**
Modal Verbs	82.24	**88.88**	93.67	88.10	87.10	90.00	76.58	86.65
Double Negation	81.73	87.26	93.61	88.12	87.26	89.00	77.28	86.32
Ensemble	82.26$\pm$0.8	88.61$\pm$0.29	93.96$\pm$0.59	88.81$\pm$0.07	87.64$\pm$0.82	88.60$\pm$0.60	76.23$\pm$0.93	86.59$\pm$0.17

Ablation Study

Methods	**STS12**	**STS13**	**STS14**	**STS15**	**STS16**	**STS-B**	**SICK-R**	**Avg.**
SimCSE-RoBERTa$_{base}$	70.16	81.77	73.24	81.36	80.65	80.22	68.56	76.57
Modal Verbs (MV)	**72.37**	**83.80**	**77.51**	**83.58**	**82.29**	**82.98**	**74.69**	**79.60**
w/o Negation as Negative	71.80	82.91	76.19	83.50	81.55	82.37	68.78	78.16
w/o Positive Augmentation	68.48	82.23	73.70	81.15	81.56	81.26	72.17	77.22

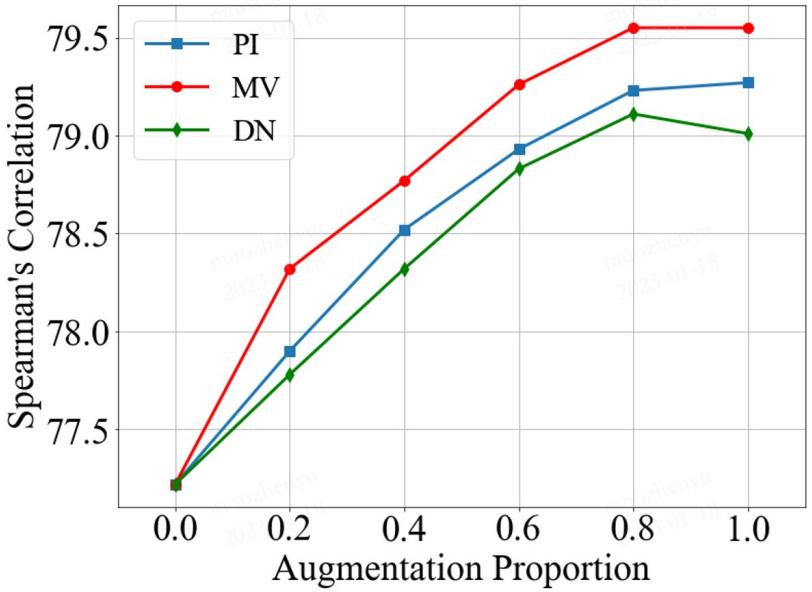
Results on Chinese Datasets

Methods	Chinese STS-B
Roberta-base (last CLS)	68.25
SimCSE-RoBERTa-base	71.10
Punctuation Insertion	71.98
Modal Verbs	**72.12**
Double Negation	71.65

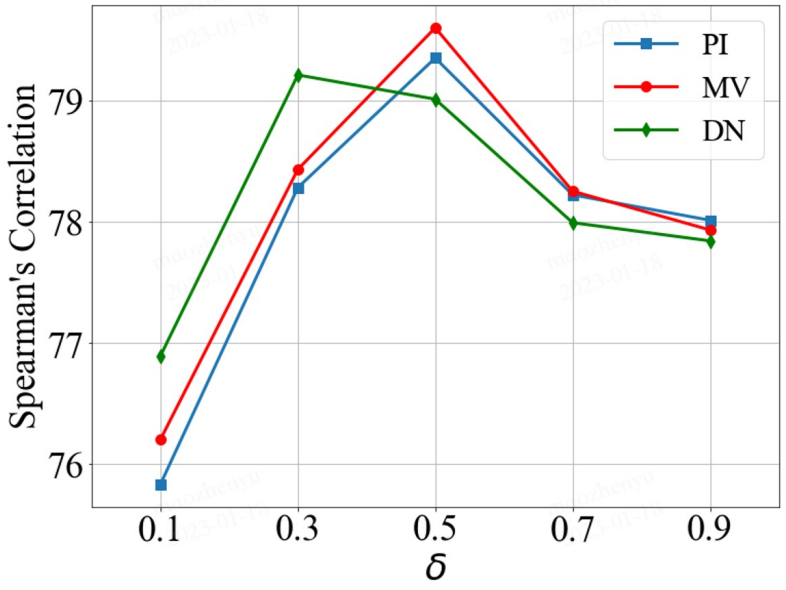
Random Manipulation VS. Rule-based Augmentation

Augmentations			**STS-B**
None (Unsup-SimCSE)			82.45
Cropping	10%	20%	30%
	77.81	71.38	63.62
Word Deletion	10%	20%	30%
	75.89	72.20	68.24
Synonym Replacement			77.45
Mask 15%			62.21
Word Repetition			84.09
Punctuation Insertion			84.55
Modal Verbs			**84.99**
Double Negation			84.12

Parameter Analysis



(a)



(b)

Thank you!