PeerDA: Data Augmentation via Modeling Peer Relation for Span Identification Tasks

香港中文大學 The Chinese University of Hong Kong

達摩院 ALIBABA DAMO ACADEMY

Weiwen Xu[1,2], Xin Li[2], Yang Deng[1], Wai Lam[1], Lidong Bing[2]

[1]The Chinese University of Hong Kong

[2]DAMO Academy, Alibaba Group

Motivation

Span identification (SpanID)

- Identify specific text spans from text input.
- Classifying the text spans into pre-defined categories.

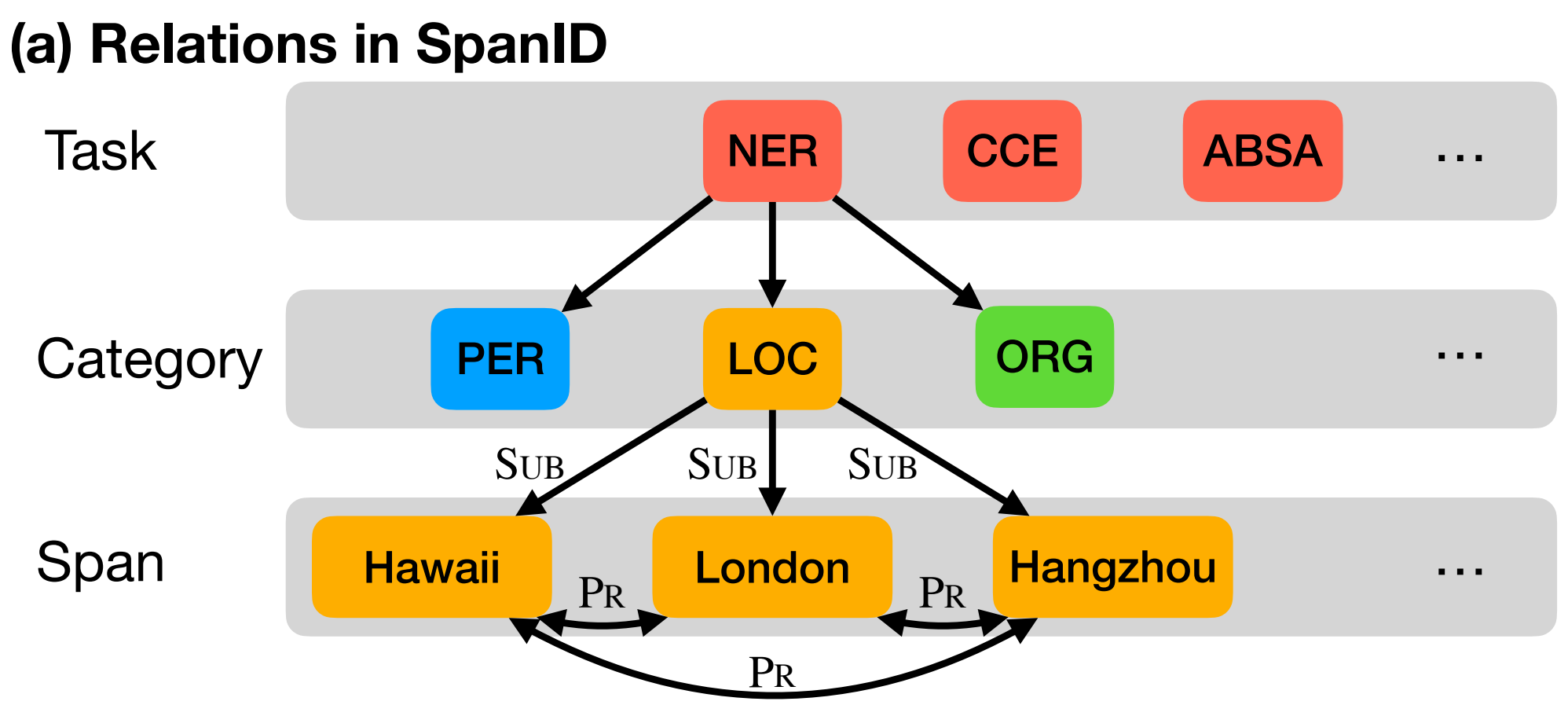
Subordinate (SUB) relation: SUB pairs $= \{(x, y) \mid x \in y\}$

- Over-fitting: Models capture superficial span-category correlations.
- Data Scarcity: SUB pairs are limited in low-resource scenarios.

Peer (PR) relation: PR pairs $= \{(x_1, x_2) \mid x_1 \in y, x_2 \in y\}$

- Jointly recognizing SUB and PR relation reduces the risk of over-fitting.
- $|$ PR pairs $| \propto |$ SUB pairs $|^2$

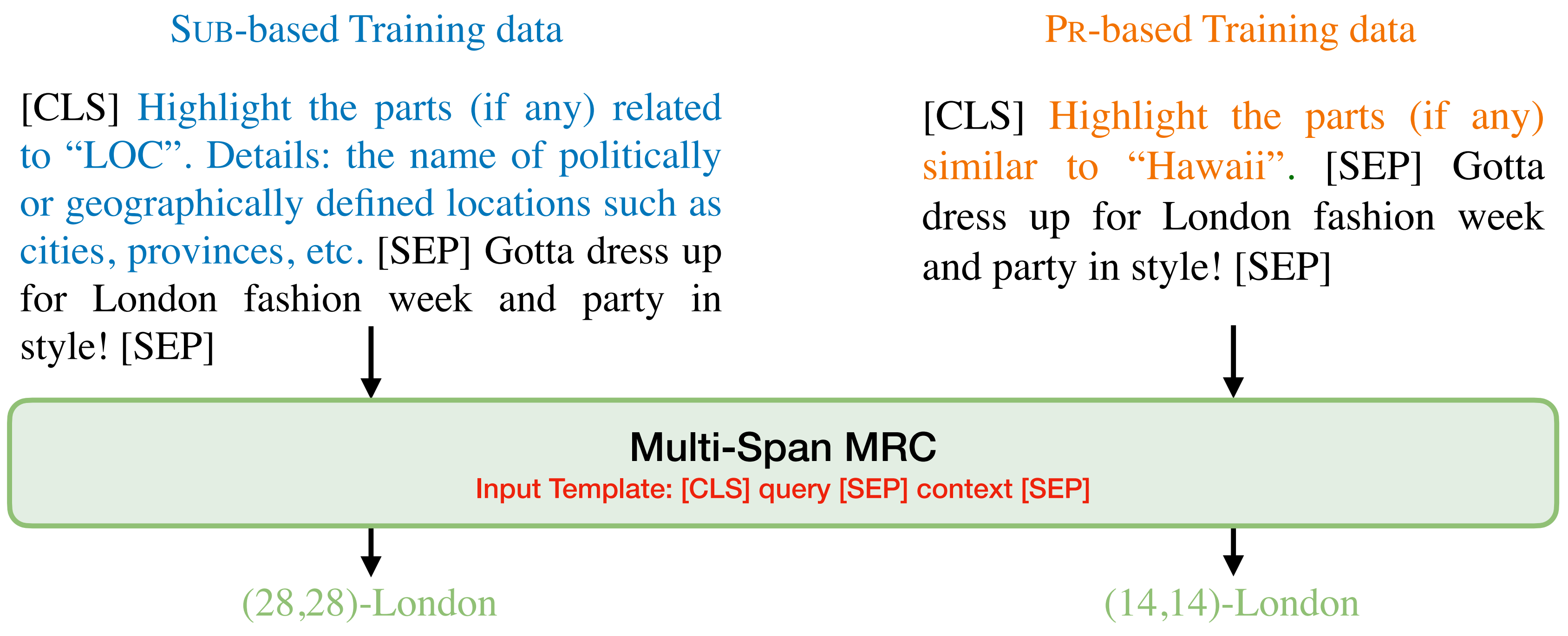
(a) Relations in SpanID



(b) SpanID in MRC Paradigm

Context:	Gotta dress up for London fashion week and party in style!	
Original data	SUB Query: Answer:	Highlight the parts (if any) related to “LOC”. Details: the name of politically or geographically defined locations such as cities, provinces, etc. London
Augmented data	PR Query-1: Answer:	Highlight the parts (if any) similar to “Hawaii”. London
	PR Query-2: Answer:	Highlight the parts (if any) similar to “Hangzhou”. London

Data Augmentation



SUB-based Training data

[CLS] Highlight the parts (if any) related to “LOC”. Details: the name of politically or geographically defined locations such as cities, provinces, etc. [SEP] Gotta dress up for London fashion week and party in style! [SEP]

PR-based Training data

[CLS] Highlight the parts (if any) similar to “Hawaii”. [SEP] Gotta dress up for London fashion week and party in style! [SEP]

Multi-Span MRC

Input Template: [CLS] query [SEP] context [SEP]

(28,28)-London

(14,14)-London

- SUB-based Query

$$Q_y^{\mathrm{SUB}} = \texttt{Highlight the parts (if any) related to } [\texttt{Men}]_y\texttt{. Details: } [\texttt{Def}]_y\texttt{.}$$

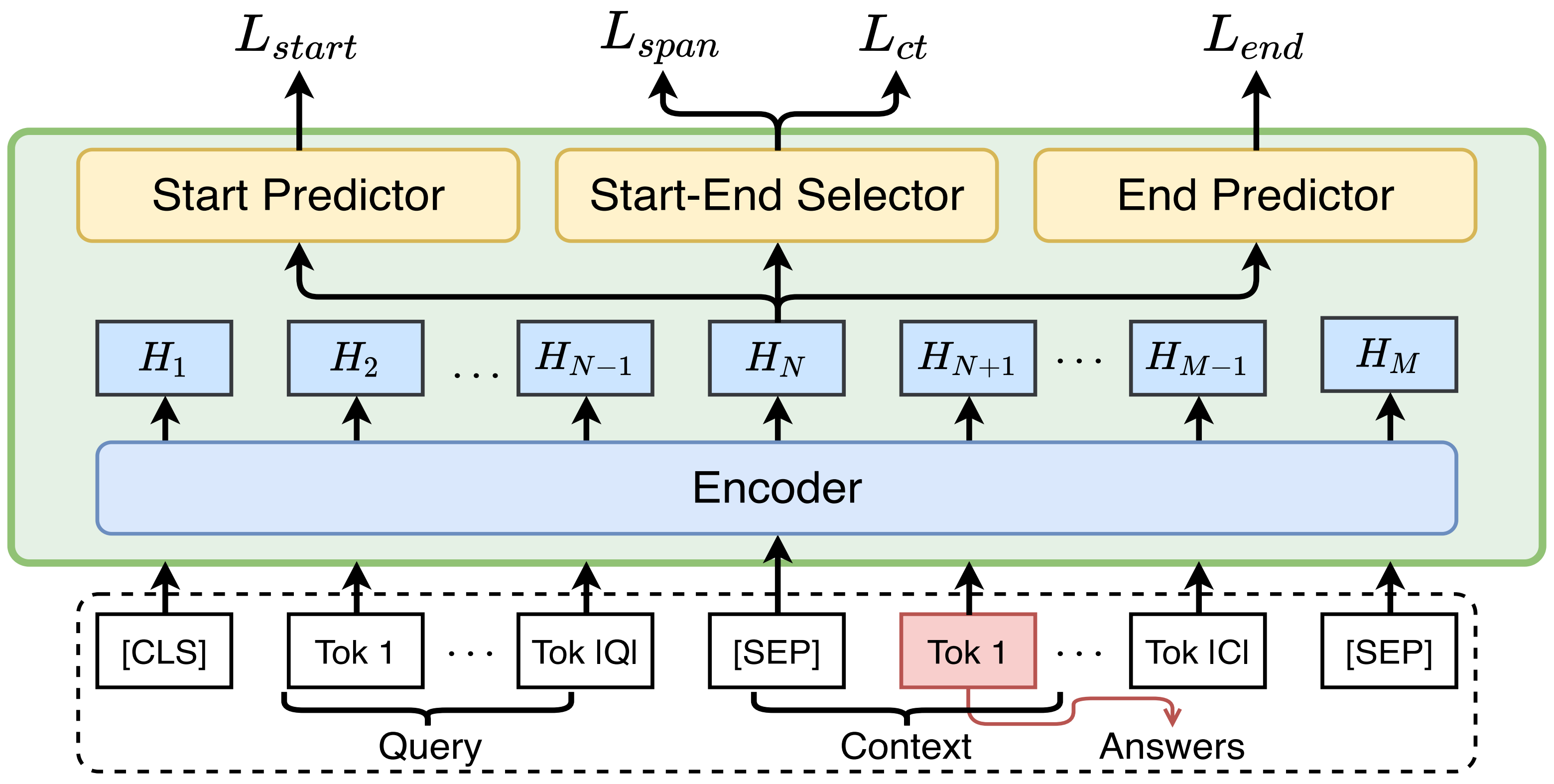
- PR-based Query

$$Q_y^{\mathrm{PR}} = \texttt{Highlight the parts (if any) similar to } x^q\texttt{.}$$

Variants

- PeerDA-Size
- PeerDA-Categ
- PeerDA-Both

Model



$$X = [[\mathrm{CLS}], Q, [\mathrm{SEP}], C, [\mathrm{SEP}]]$$
$$H = \mathrm{Encoder}(X)$$
$$P_{end} = HW^e, \quad P_{start} = HW^s$$
$$P_{span} = \boldsymbol{FFN}(H)^T H$$

$$L_{start} = \mathbf{CE}(\sigma(P_{start}), Y_{start})$$
$$L_{end} = \mathbf{CE}(\sigma(P_{end}), Y_{end})$$
$$L_{span} = \mathbf{CE}(\sigma(P_{span}), Y_{span})$$
$$L_{cl} = \mathbf{CL}(\sigma(P_{span}), \sigma(P_{span'}))$$

SpanID Results

Methods	OntoNotes5			WNUT17			Movie			Restaurant			Weibo		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
	RB-CRF+RM			CL-KL			T-NER			KaNa			RoBERTa+BS		
SOTA	92.8	92.4	92.6	-	-	**60.5**	-	-	71.2	80.9	80.0	80.4	70.2	75.4	**72.7**
							Base								
Tagging	91.0	91.8	91.4	62.1	48.2	54.3	73.0	72.8	72.9	80.6	80.7	80.7	70.8	71.0	70.9
MRC	92.4	91.8	92.1	66.4	40.7	50.5	70.3	73.3	71.8	81.4	79.9	80.6	73.6	64.4	68.7
PeerDA	91.9	92.6	92.4	71.1	46.9	56.5	77.9	72.3	75.0	81.3	82.8	82.1	70.0	73.3	71.6
							Large								
Tagging	93.0	92.3	92.6	69.4	46.2	55.4	74.2	74.0	74.1	80.9	82.0	81.4	71.4	69.2	70.3
MRC	92.8	91.8	92.3	72.4	41.7	52.9	76.7	73.2	74.9	81.6	81.7	81.7	72.2	66.8	69.4
PeerDA	92.8	93.7	**93.3**	70.9	48.0	57.2	78.5	73.1	**75.7**	81.8	82.5	**82.2**	73.4	71.6	72.5

NER

Methods	Lap14		Rest14	
	UABSA	ATE	UABSA	ATE
SPAN-BERT	61.3	82.3	73.7	86.7
IMN-BERT	61.7	77.6	70.7	84.1
RACL	63.4	81.8	75.4	86.4
Dual-MRC	65.9	82.5	**76.0**	86.6
MRC (Large)	63.2	83.9	72.9	86.8
PeerDA	**65.9**	**84.6**	73.9	**86.8**

ABSA

Methods	#Params	AUPR	P@0.8R
ALBERT$_{\texttt{xxlarge}}$	223M	38.4	31.0
RoBERTa$_{\texttt{base}}$ + CP	125M	45.2	34.1
RoBERTa$_{\texttt{large}}$	355M	48.2	38.1
DeBERTa$_{\texttt{xlarge}}$	900M	47.8	44.0
ConReader$_{\texttt{large}}$	355M	49.1	44.2
MRC (Base)	125M	43.6	32.2
PeerDA	125M	**52.3**	**45.5**

CCE

Methods	News20			Social21		
	P	R	F1	P	R	F1
Volta	-	-	-	50.1	46.4	48.2
HOMADOS	-	-	-	41.2	40.3	40.7
TeamFPAI	-	-	-	65.2	28.6	39.7
MRC (Base)	10.5	53.5	17.6	55.8	43.5	48.9
PeerDA	21.8	31.5	**25.8**	49.4	70.6	**58.1**

SBPD

Ablation

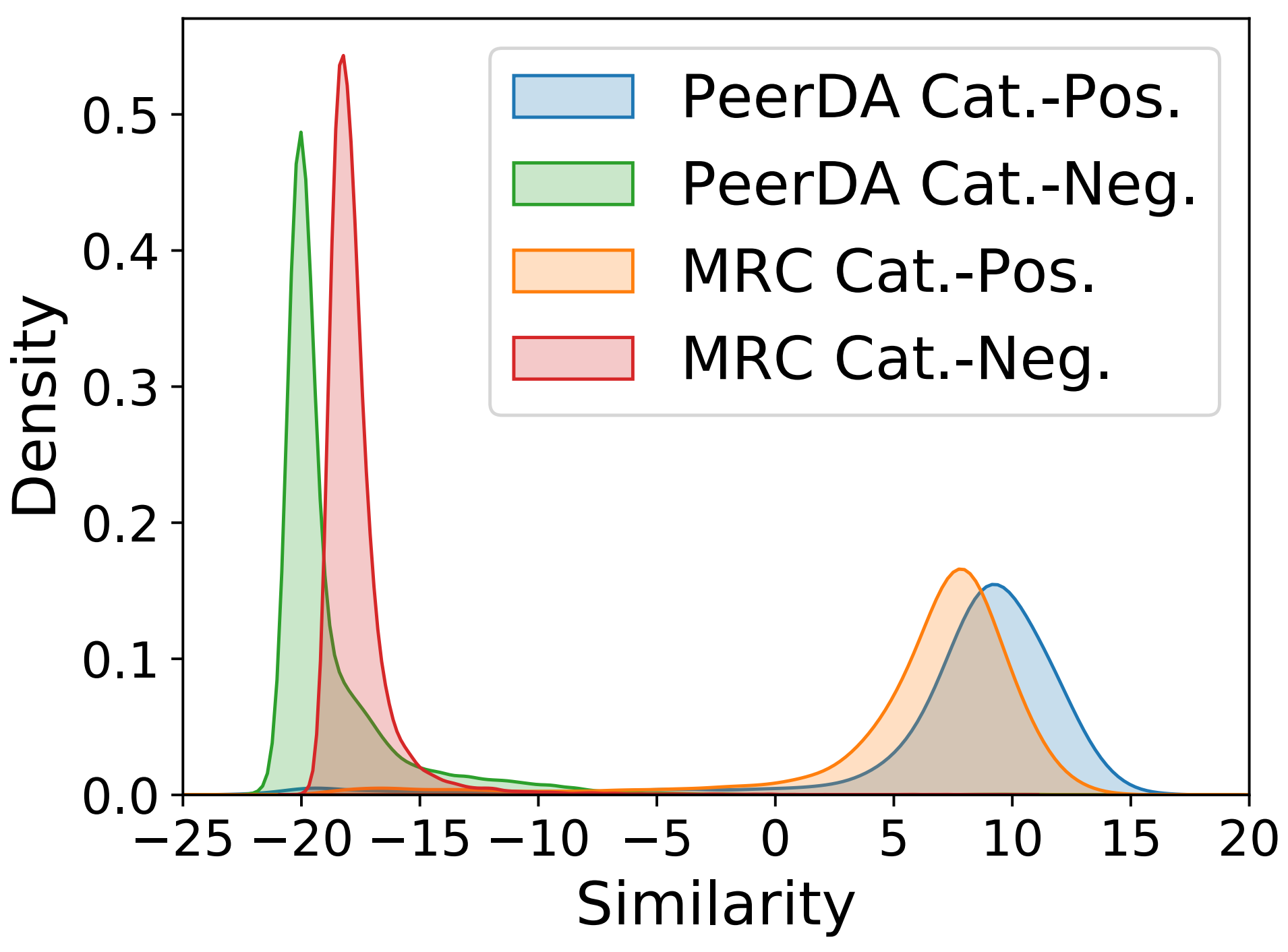
Ablation Type	NER	UABSA	SBPD	CCE	Avg.
MRC	72.7	68.1	33.3	43.6	54.4
PeerDA-Size	74.6	69.7	38.5	48.7	57.9
PeerDA-Categ	74.2	69.3	40.4	51.3	58.8
PeerDA-Both (**final**)	**75.5**	**69.9**	**42.0**	**52.3**	**59.9**

DA Strategy

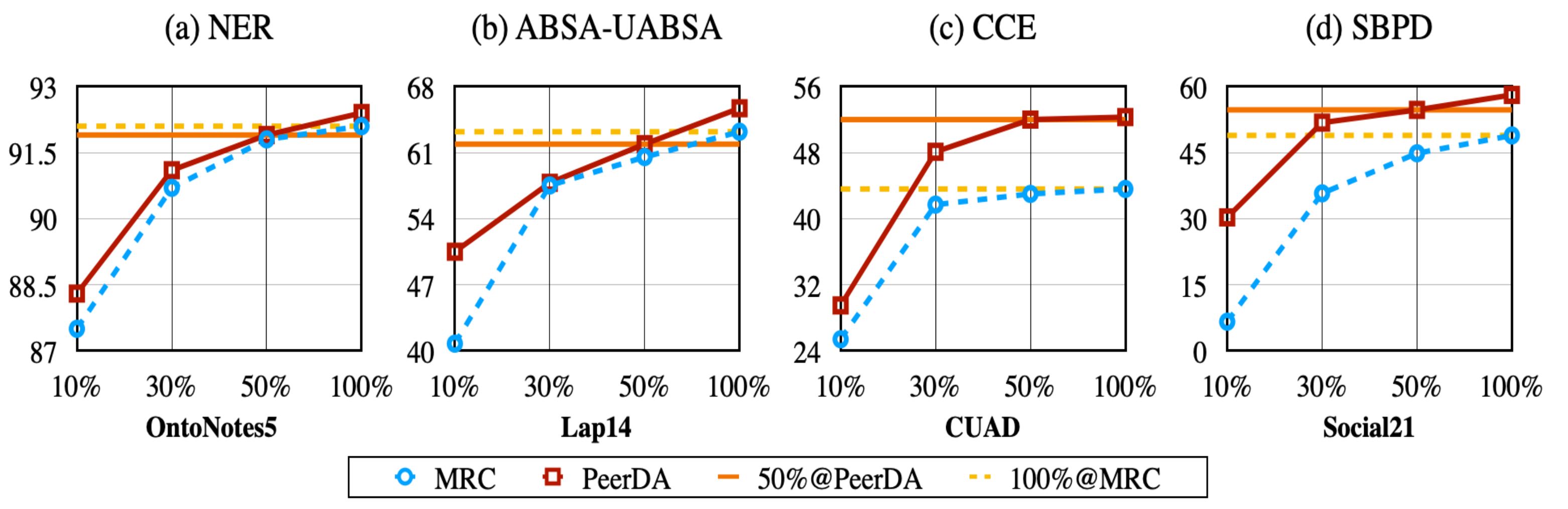
Ablation Type	|GPU|	NER	UABSA	SBPD	Avg.
Calculation of $P_{s,e}$					
concat	1x	74.5	69.2	40.3	61.3
general (**final**)	**0.23x**	**75.0**	**69.4**	**40.8**	**61.7**
Contrastive Loss					
Average	0.23x	75.1	69.6	37.6	60.8
Max-Min (**final**)	0.23x	**75.5**	**69.9**	**42.0**	**62.4**

Model Design

Semantic Distance



Low-resource



Reproducibility

Codes are available at https://github.com/DAMO-NLP-SG/PeerDA