Curriculum-Meta Learning for Order-Robust Continual Relation Extraction

Authors:

Tongtong Wu, Xuekai Li, Yuan-Fang Li, Reza Haffari, Guilin Qi, Yujin Zhu, Guoqiang Xu

Repoter: Xuekai Li

SOUTHEAST UNIVERSITY

MONASH University

The Next

Part 1

Introduction

1.1 Background

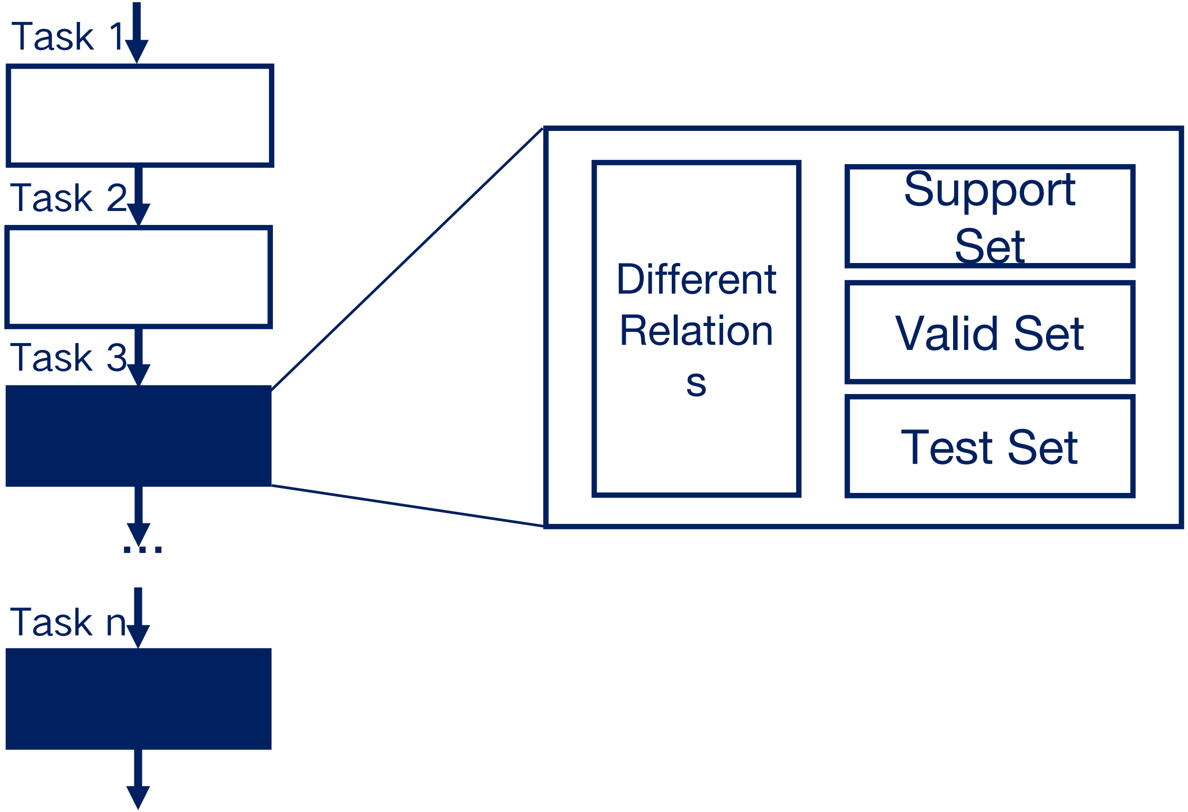
Relation Extraction

unstructured text

structured facts triples

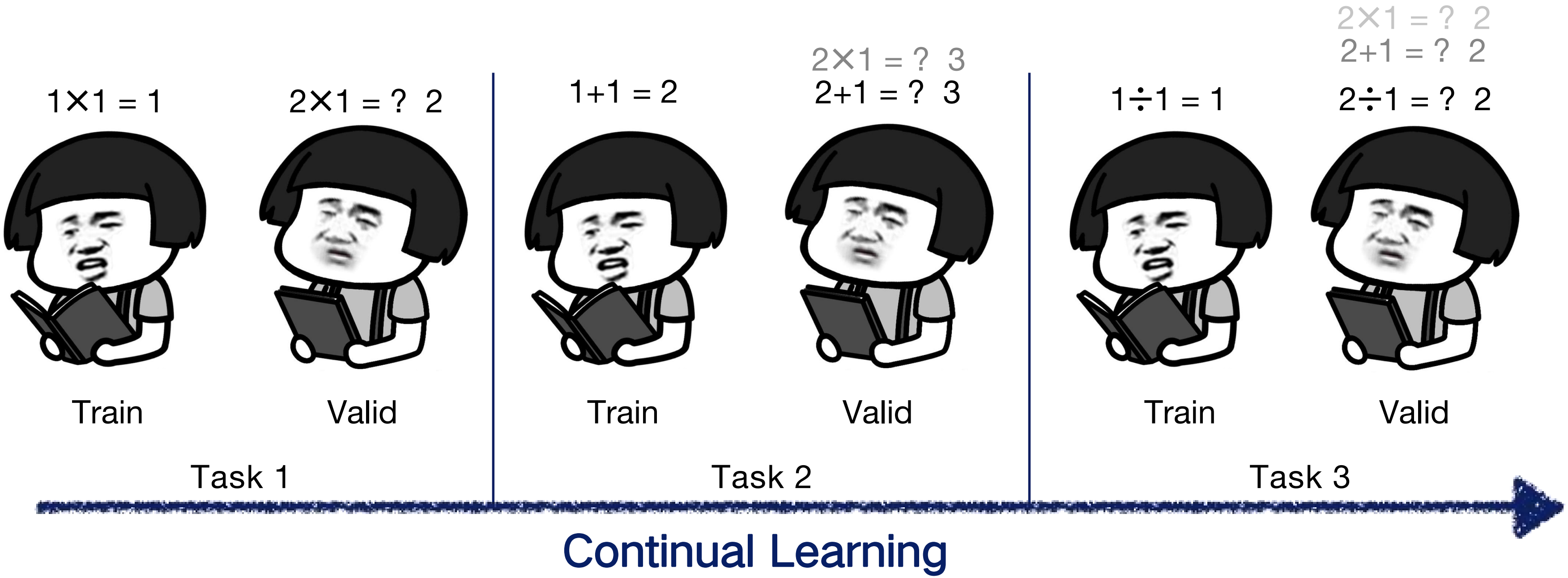
1.1 Background

Continual Relation Extraction



1.2 Challenge

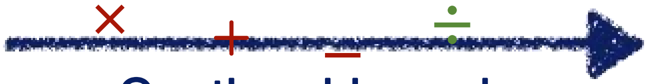
Catastrophic Forgetting



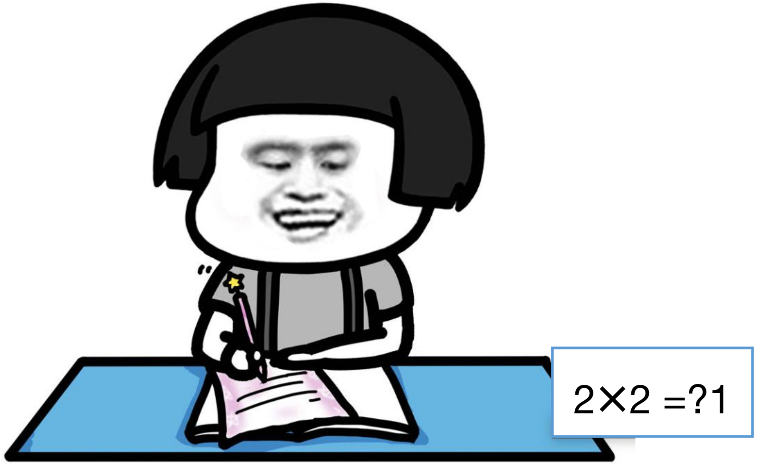
1.2 Challenge

Catastrophic Forgetting

When a neural network is utilized to learn a sequence of tasks, the learning of the later tasks may degrade the performance of the learned model for the previous tasks.



Continual Learning



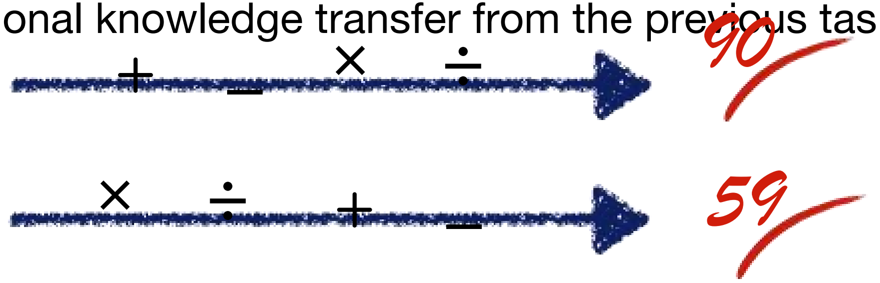
Test

1.2 Challenge

Order-sensitivity

The performance of the tasks vary based on the order of the task arrival sequence.

- CF incurred by the different sequences of previous tasks
- the unidirectional knowledge transfer from the previous tasks.



(i) Ethical AI considerations in continual learning, e.g. fairness in the medical domain [Yoon et al. 2020];
(ii) Benchmarking of continual learning algorithms as most of the existing works pick an arbitrary and random sequence of the given tasks for the evaluation [Chen et al. 2018];
(iii) Uncertainty to the quality of extracted knowledge in the realistic scenario for knowledge base population, where the model is faced with only one sequence.

1.3 Our Contribution

Three Contributions:

- A novel curriculum-based continual learning method tackling the order-sensitivity and catastrophic forgetting problems in continual relation extraction.
- A new relation representation learning method via the conceptual distribution of domain and range of relations
- Comprehensive experiments to analyze the order-sensitivity and catastrophic forgetting problems in state-of-the-art models

The Next

Part 2

Related Works

2 Related Works

Continual Learning

GEM [Lopez-Paz et al. 2017]   Experience Replay-based Model

EWC [Kirkpatrick et al. 2016]   Weight Consolidation-based Model

- R-EWC [Liu et al. 2016]

Continual Relation Extraction

EA-EMR [Wang et al. 2019]

MLLRE [Obamuyide et al. 2019]

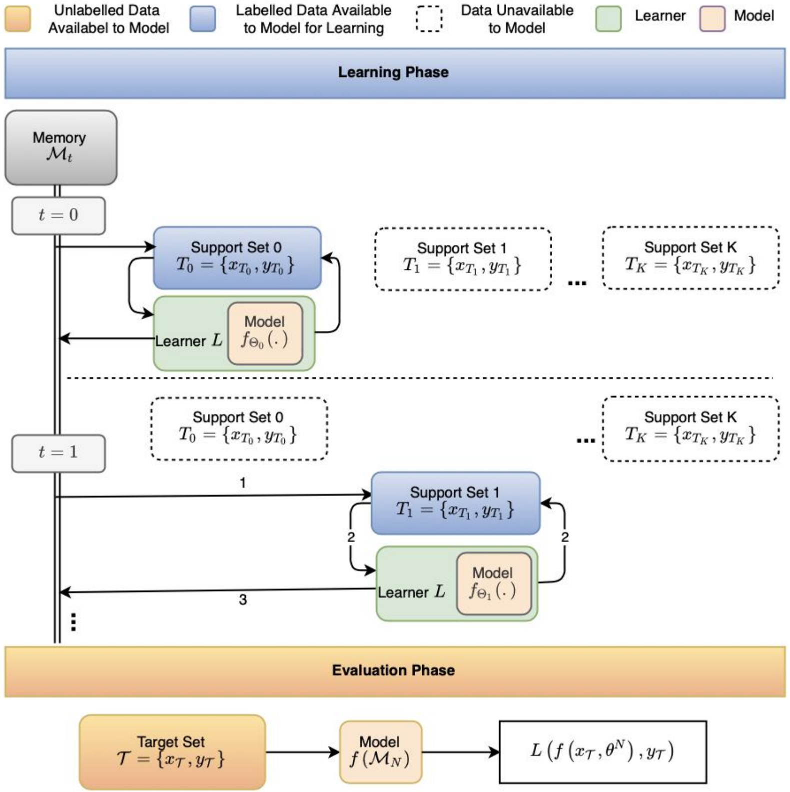
EMAR [Han et al. 2020]

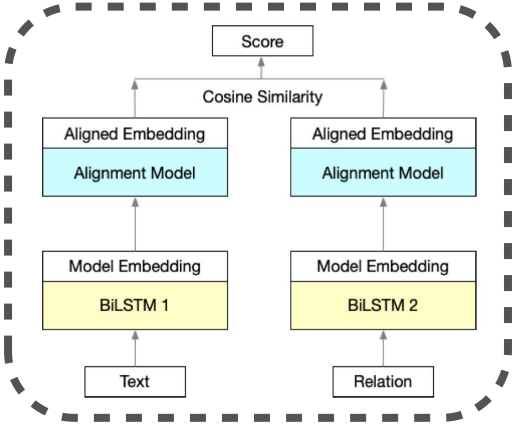
The Next

Part 3

Curriculum-meta Learning

3.1 Continual Relation Extraction

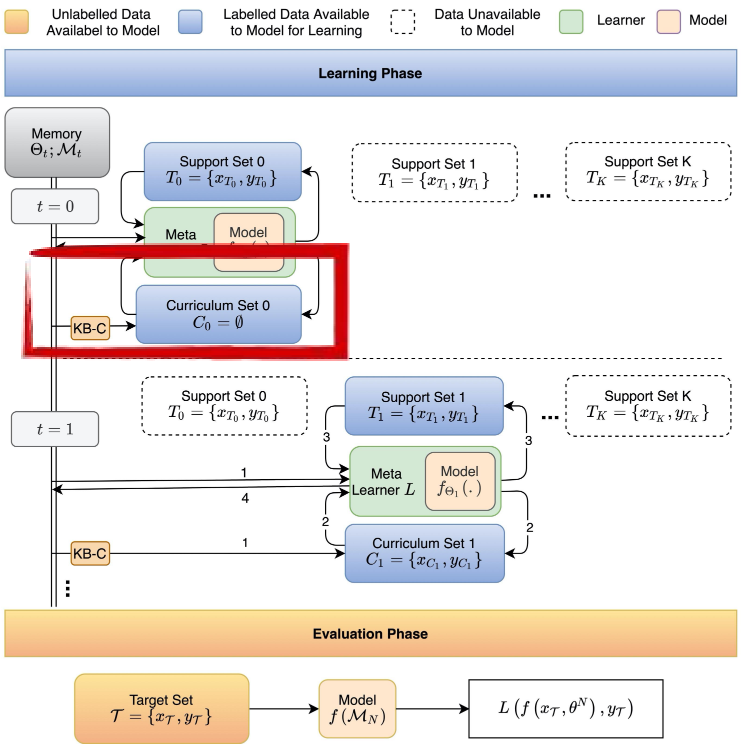




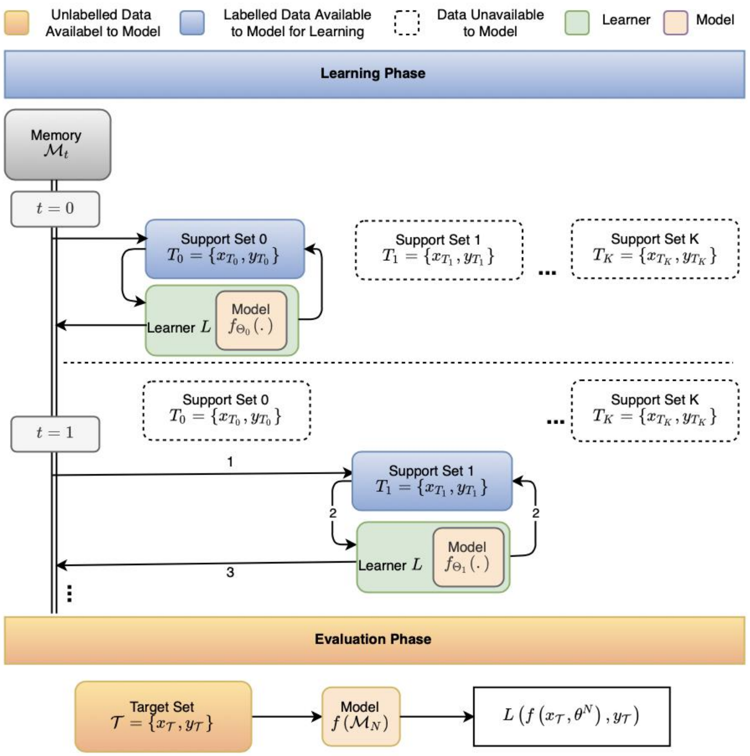
Vanilla Model

$$l = max\{0, \gamma - sim(r^+, x) + sim(r^-, x)\}$$

3.2 Framework

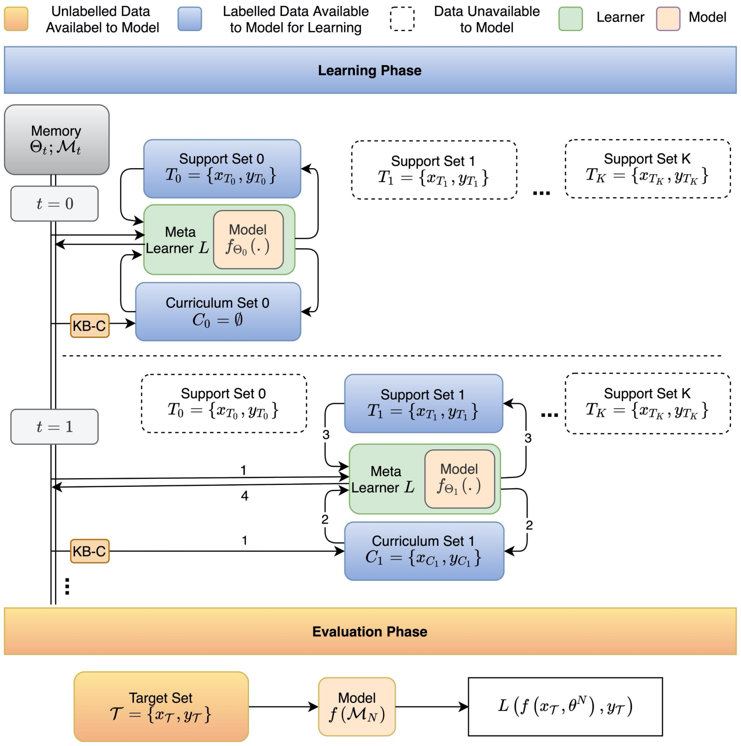


Our Framework



EA - EMR

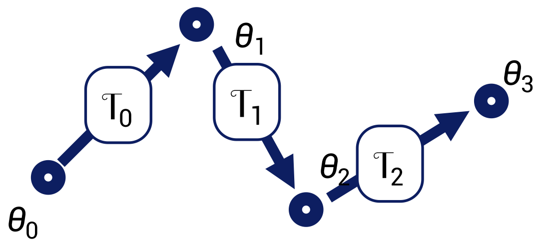
3.3 Curriculum-Meta Learning



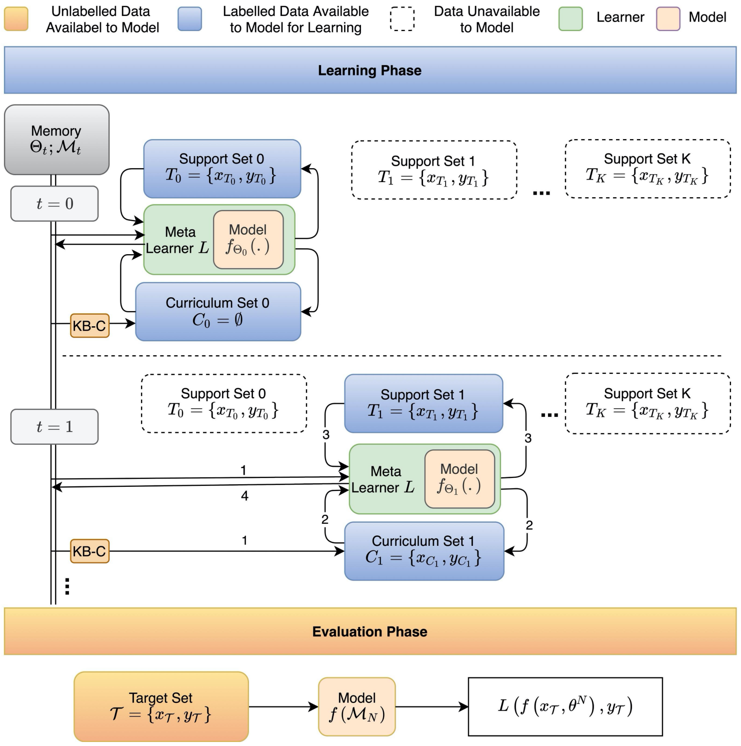
CML

Meta-Training

Conventional Machine Learning:



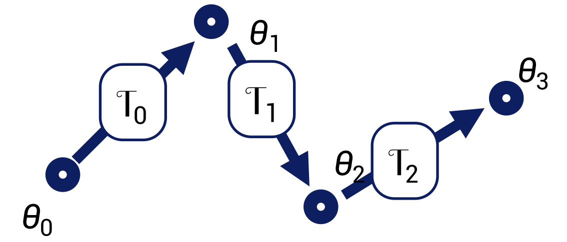
3.3 Curriculum-Meta Learning



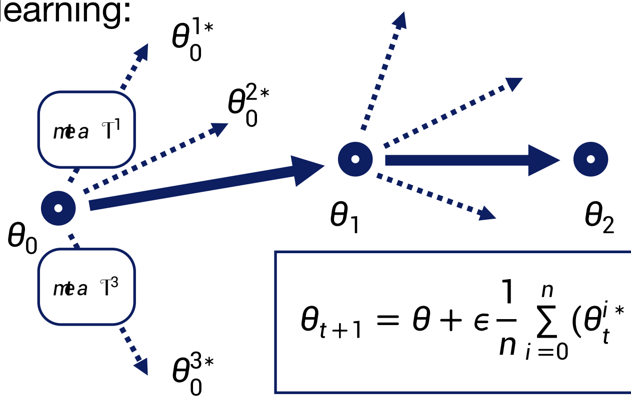
CML

Meta-Training

Conventional Machine Learning:

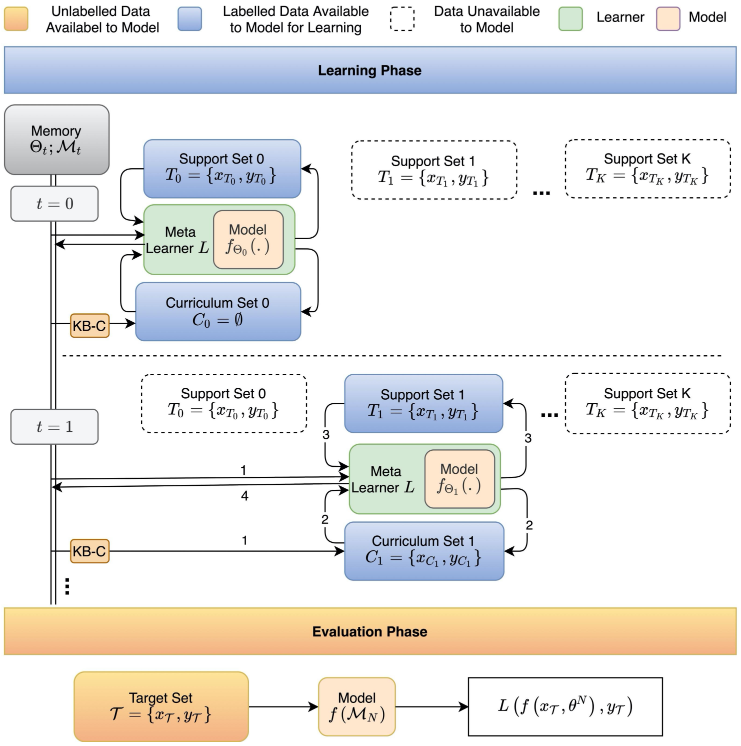


Meta-learning:



$$\theta_{t+1} = \theta + \epsilon \frac{1}{n} \sum_{i=0}^{n} (\theta_t^{i*} - \theta_t)$$

3.3 Curriculum-Meta Learning



CML

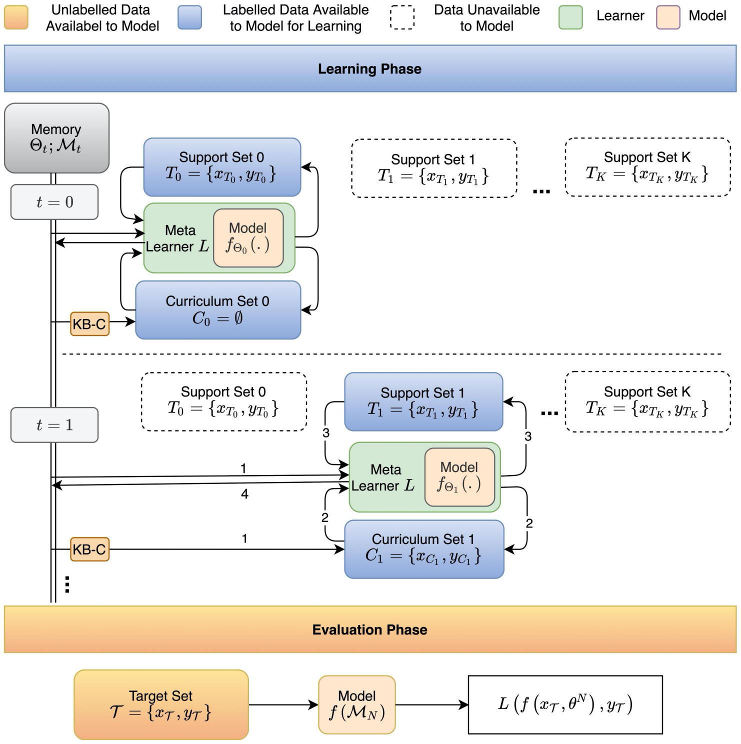
Curriculum-based Memory Replay

1. Assessing the difficulty of tasks.

2. Sampling instances from the memory.

3. Ranking the sampled instances by a certain strategy, inducting the model to learn the bias between the current task and observed similar tasks.

3.4 Knowledge-based Curriculum

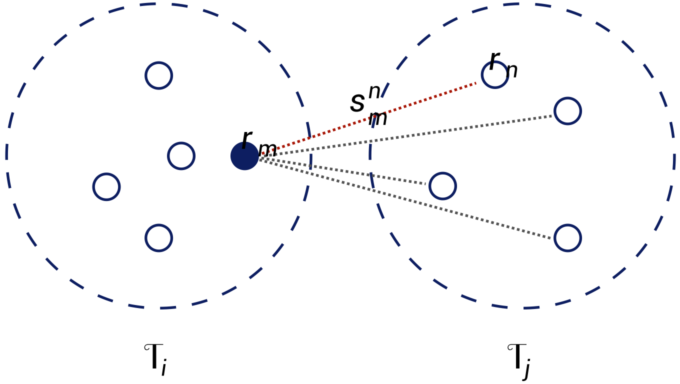


CML

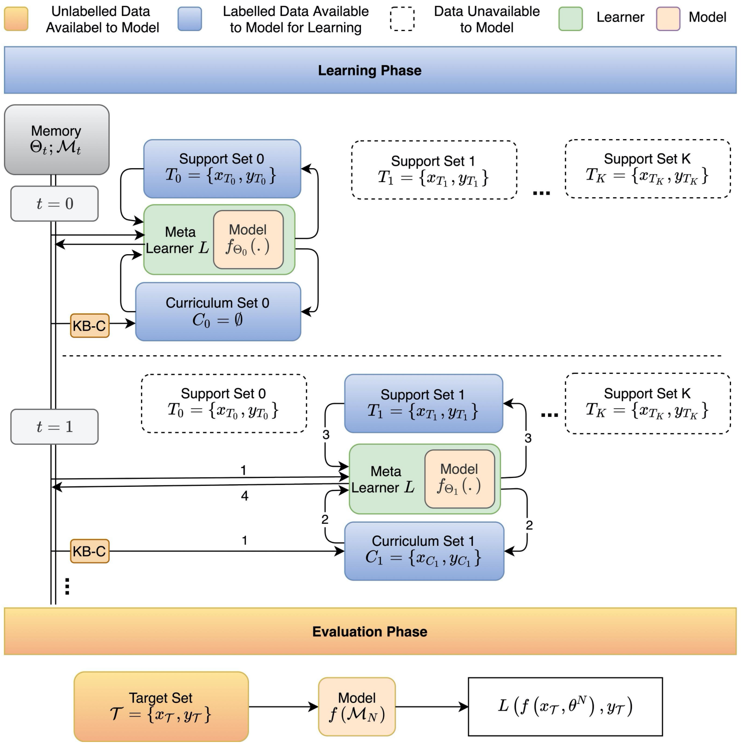
KB-C

Difficulty Function:

$$DI_i := \frac{1}{K-1} \sum_{j=1; j \neq i}^{K} S_i^j \qquad S_i^j = \frac{1}{M \times N} \sum_{m=1}^{M} \sum_{n=1}^{N} s_m^n$$



3.4 Knowledge-based Curriculum

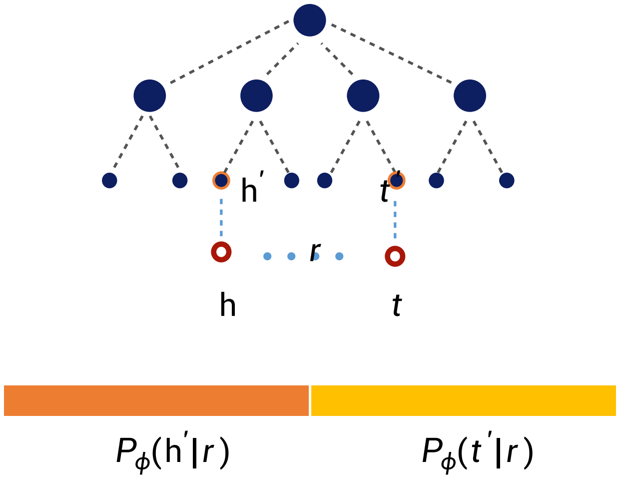


CML

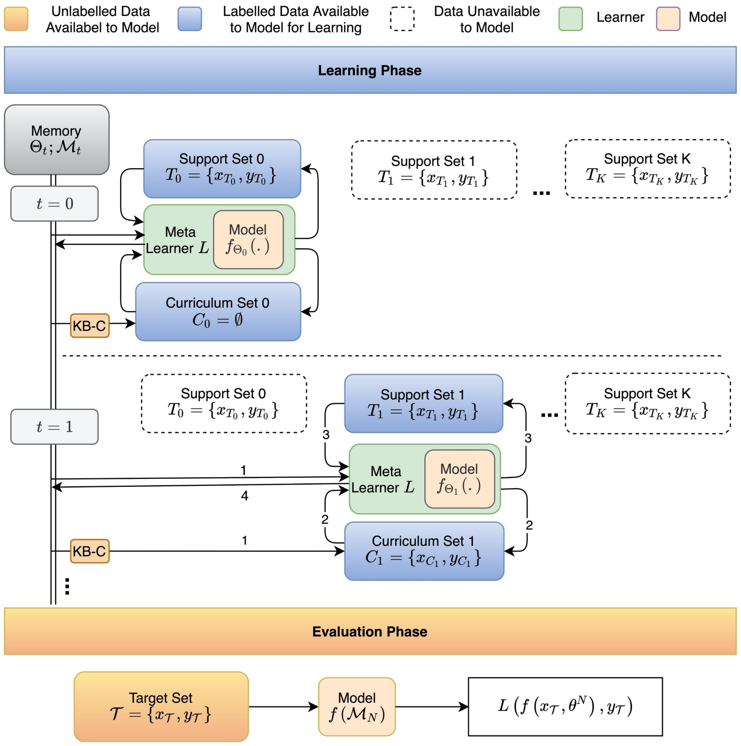
KB-C

Representation Function:

$$\min_{\phi} \mathcal{L}(\phi, \mathcal{G}) = \min \sum_{(h',t';r)\in\mathcal{G}} [-\log P_\phi(h'|r) - \log P_\phi(t'|r)]$$



3.4 Knowledge-based Curriculum

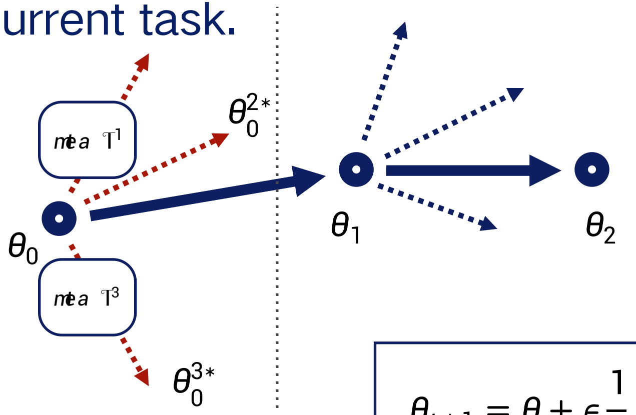


CML

KB-C

Sampling Strategy:

Select and sort memory-stored instances of the most similar relations to current task.



$$\theta_{t+1} = \theta + \epsilon \frac{1}{n} \sum_{i=0}^{n} (\theta_t^{i*} - \theta_t)$$

The Next

Part 4

Experiments and Discussi

4.1 Experiment Settings

Baselines

Vanilla Model [Yu et al. 2017]

EA-EMR [Wang et al. 2019]

MLLRE [Obamuyide et al. 2019]

EMAR [Han et al. 2020]

Benchmarks [Wang et al. 2019]

Lifelong - Fewrel: 80 relations and 700 instances per relation.

Lifelong - SimpleQuestions: 1,785 relations and totally 72,238 instances.

Lifelong - Tacred: 42 relations and totally 21,784 instances.

4.2 Main Results

The average accuracy Acc_a and whole accuracy Acc_w, with error bounds EB, on the test sets of observed tasks at the final time step.

Metrics:

$$Acc_w := acc_{f,\mathfrak{D}_{test}}$$

$$Acc_a := \frac{1}{k}\sum_{i=1}^{k} acc_{f,\mathfrak{D}_{test}^{i}}$$

$$EB := Z_{\frac{\alpha}{2}} \times \frac{\delta}{\sqrt{n}}$$

		Continual-FewRel				Continual-SimpQ				Continual-TACRED			
		Acc_w		Acc_a		Acc_w		Acc_a		Acc_w		Acc_a	
Setting	Model	Acc	EB	Acc	EB	Acc	EB	Acc	EB	Acc	EB	Acc	EB
Cluster	Vanilla‡	16.3	±4.10	19.7	±3.90	60.3	±2.52	58.3	±2.30	12.0	±3.21	8.7	±2.35
	EWC†	27.1	±2.32	30.2	±2.10	67.2	±3.16	59.0	±2.20	14.5	±2.51	14.5	±2.90
	AGEM†	36.1	±2.51	42.5	±2.63	77.6	±2.11	72.2	±2.72	12.5	±2.24	16.5	±2.20
	EA-EMR‡	59.8	±1.50	74.8	±1.30	82.7	±0.48	86.2	±0.33	17.8	±1.01	25.4	±1.17
	EMAR†	53.8	±1.30	68.6	±0.71	80.0	±0.83	76.9	±1.39	42.7	±2.92	**52.5**	±1.74
	MLLRE	56.8	±1.30	70.2	±0.93	84.5	±0.35	86.7	±0.46	34.4	**±0.49**	41.2	±1.37
	CML (ours)	**60.2**	**±0.71**	**76.0**	**±0.24**	**85.6**	**±0.34**	**87.5**	**±0.32**	**44.4**	±1.16	49.3	**±1.01**
Random	Vanilla‡	19.1	±1.20	19.3	±1.30	55.0	±1.30	55.2	±1.30	10.2	±2.02	10.4	±2.31
	EWC†	30.1	±1.07	30.2	±1.05	66.4	±0.81	66.7	±0.83	15.3	±1.70	15.4	±1.79
	AGEM†	36.9	±0.80	37.0	±0.83	76.4	±1.02	76.7	±1.01	13.4	±1.47	14.3	±1.62
	EA-EMR‡	61.4	±0.81	61.6	±0.76	83.1	±0.41	83.2	±0.47	27.3	±1.01	30.3	±0.70
	EMAR†	62.7	±0.63	62.8	±0.62	82.4	±0.86	84.0	±0.78	**45.1**	±1.48	**46.4**	±2.00
	MLLRE	59.8	±0.91	59.8	±0.94	85.2	±0.25	85.5	±0.31	36.4	**±0.66**	38.0	**±0.58**
	CML (ours)	**62.9**	**±0.62**	**63.0**	**±0.59**	**86.5**	**±0.22**	**86.9**	**±0.28**	43.7	±0.83	45.3	±0.72

4.3 Case Study

A case study of EA-EMR, MLLRE and Our CML on the FewRel dataset.

$taskID$	T_0	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	Acc_a	Acc_w
$runID$	0	1	2	3	4	5	6	7	8	9		
P_0	93.3	36.7	61.7	57.7	100.0	26.5	35.7	58.1	54.1	50.6	76.0	60.9
P_1	91.9	42.4	62.5	64.6	99.3	24.5	60.3	56.2	62.2	61.8	74.0	61.7
P_2	92.6	42.7	60.6	66.1	100.0	25.8	55.4	69.3	71.1	54.5	73.6	61.3
P_3	93.3	39.2	68.4	60.2	100.0	34.0	57.3	71.4	85.2	58.3	74.9	63.4
P_4	89.6	34.4	64.0	59.9	100.0	35.1	65.8	71.0	77.8	72.5	75.6	55.8
P_5	91.1	60.9	56.9	64.2	99.3	45.7	77.9	77.1	67.4	79.9	75.1	58.3
P_6	94.1	66.9	72.1	55.5	100.0	53.0	77.2	74.3	84.4	74.0	74.5	59.2
P_7	94.8	73.6	85.4	53.6	98.6	61.6	80.4	81.5	92.6	86.5	76.0	62.0
P_8	94.8	74.6	85.2	85.0	98.6	69.1	79.3	75.2	87.4	92.9	74.2	54.5
P_9	91.1	76.9	91.3	89.4	97.9	65.4	77.6	83.9	93.3	85.8	76.1	58.9
μ	92.7	54.8	70.8	65.6	99.4	44.1	66.7	71.8	77.6	71.7	75.00	59.60
δ	1.73	17.34	12.22	12.09	0.77	17.25	14.61	8.97	13.36	14.75	0.91	2.83

CML

$taskID$	T_0	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	Acc_a	Acc_w
$runID$	0	1	2	3	4	5	6	7	8	9		
P0	88.1	39	49.9	59.9	100	23.8	28.2	48.5	52.6	53.9	76.8	62.7
P1	82.2	43.5	68.8	33.9	100	30.7	56.7	39.6	71.1	59	73.4	63.9
P2	83	49.9	66.7	75.9	100	26.7	57.4	74.7	58.5	64.1	72.6	61.2
P3	91.9	48.3	76.3	77	100	23.2	54.3	77	83.7	50.6	62.8	59
P4	90.4	44.1	73.6	79.6	100	43.5	47.5	69.6	86.7	77.6	77.7	58.4
P5	97	42.8	71.2	79.9	100	46	74.9	68.5	80	76.8	79.5	61.6
P6	98.5	79.2	69.3	75.5	100	52.4	82.6	88.1	88.1	80.4	74.6	58.7
P7	97	80.2	92.1	67.9	100	57.1	81.9	87.6	93.3	81.4	67.4	56.3
P8	91.9	81.5	91.9	92.7	99.3	68	86.1	93.4	92.6	90.8	77.4	56.1
P9	100	89.7	97.4	96	100	82.1	91.9	94.7	99.3	93.1	77.7	58.7
μ	92	59.8	75.7	73.8	99.9	45.4	66.2	74.2	80.6	72.8	74	59.7
δ	6.25	20.06	14.39	17.51	0.22	19.92	20.41	18.50	15.35	14.99	5.26	2.61

EA-EMR

Summary: CML alleviate the order-sensitivity problem.

4.3 Case Study

A case study of EA-EMR, MLLRE and Our CML on the FewRel dataset.

$taskID$	T_0	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	Acc_a	Acc_w
$runID$	0	1	2	3	4	5	6	7	8	9		
P_0	93.3	36.7	61.7	57.7	100.0	26.5	35.7	58.1	54.1	50.6	76.0	60.9
P_1	91.9	42.4	62.5	64.6	99.3	24.5	60.3	56.2	62.2	61.8	74.0	61.7
P_2	92.6	42.7	60.6	66.1	100.0	25.8	55.4	69.3	71.1	54.5	73.6	61.3
P_3	93.3	39.2	68.4	60.2	100.0	34.0	57.3	71.4	85.2	58.3	74.9	63.4
P_4	89.6	34.4	64.0	59.9	100.0	35.1	65.8	71.0	77.8	72.5	75.6	55.8
P_5	91.1	60.9	56.9	64.2	99.3	45.7	77.9	77.1	67.4	79.9	75.1	58.3
P_6	94.1	66.9	72.1	55.5	100.0	53.0	77.2	74.3	84.4	74.0	74.5	59.2
P_7	94.8	73.6	85.4	53.6	98.6	61.6	80.4	81.5	92.6	86.5	76.0	62.0
P_8	94.8	74.6	85.2	85.0	98.6	69.1	79.3	75.2	87.4	92.9	74.2	54.5
P_9	91.1	76.9	91.3	89.4	97.9	65.4	77.6	83.9	93.3	85.8	76.1	58.9
μ	92.7	54.8	70.8	65.6	99.4	44.1	66.7	71.8	77.6	71.7	75.00	59.60
δ	1.73	17.34	12.22	12.09	0.77	17.25	14.61	8.97	13.36	14.75	0.91	2.83

CML

$taskID$	T_0	T_1	T_6	T_3	T_4	T_5	T_6	T_7	T_8	T_9	Acc_a	Acc_w
$runID$	0	1	2	3	4	5	6	7	8	9		
P_0	83.0	44.1	48.8	53.6	100.0	25.9	49.8	52.7	63.0	54.7	74.8	59.7
P_1	91.9	39.0	51.0	62.4	100.0	25.6	61.4	60.7	48.1	53.9	75.3	60.3
P_2	94.1	43.5	62.3	70.4	100.0	23.8	56.2	70.0	72.6	44.0	71.0	60.9
P_3	84.4	49.7	72.8	62.4	100.0	31.1	58.9	71.5	80.0	62.6	72.4	62.6
P_4	97.0	46.1	62.3	80.7	100.0	31.7	63.2	68.0	70.4	78.1	75.5	56.5
P_5	95.6	53.3	59.5	75.5	100.0	45.2	68.0	80.4	70.4	72.5	76.1	57.7
P_6	95.6	72.0	71.2	59.9	99.3	45.3	77.7	77.2	84.4	70.7	74.7	57.6
P_7	95.6	76.4	89.8	59.1	98.6	52.4	81.1	80.7	92.6	70.5	73.6	59.7
P_8	91.9	77.6	85.6	88.0	97.2	68.5	77.1	81.2	85.9	86.8	74.9	56.2
P_9	96.3	73.9	92.6	88.3	94.4	70.6	78.7	84.1	93.3	88.0	73.2	55.6
μ	92.5	57.6	69.6	70.0	99.0	42.0	67.2	72.7	76.1	68.2	74.15	58.70
δ	4.97	15.52	15.61	12.51	1.84	17.43	10.94	10.11	14.07	14.41	1.58	2.30

MLLRE

Summary: CML do alleviate the order-sensitivity problem.

4.4 Alleviate Overfitting on Memory

Memory-Training Rate

Train		100		200	all
Memory		25	50	50	50
EA-EMR	Acc_a	70.7	75.5	74.8	73.9
	Acc_w	53.2	57.4	59.8	59.6
MLLRE	Acc_a	68.4	72.1	70.2	51.0
	Acc_w	51.9	57.8	56.8	47.3
EMAR	Acc_a	60.1	66.7	68.6	74.1
	Acc_w	43.7	51.2	53.8	57.7
CML	Acc_a	**73.6**	**76.4**	**76.0**	58.0
	Acc_w	**54.7**	**60.3**	**60.2**	49.1

Summary: The performance of CML is much better than MLLRE.

4.5 The effectiveness of KB-C

A New Metric

Average Forgetting Rate:

$$Fr^{j}_{avg} := \frac{1}{K-1}\sum_{i=1}^{K-1}\frac{\overline{acc}^{j}_{i+1}-\overline{acc}^{j}_{i}}{\overline{acc}^{j}_{i}}$$

$$\overline{acc}^{j}_{i} := \frac{1}{(J-1)!}\sum_{\pi\in\Pi_{[1,\dots,J]}\, st\, \pi_i=j} acc_i(\pi)$$

$taskID$	T_0	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	Acc_a	Acc_w
$runID$	0	1	2	3	4	5	6	7	8	9		
P_0	93.3	36.7	61.7	57.7	100.0	26.5	35.7	58.1	54.1	50.6	76.0	60.9
P_1	91.9	42.4	62.5	64.6	99.3	24.5	60.3	56.2	62.2	61.8	74.0	61.7
P_2	92.6	42.7	60.6	66.1	100.0	25.8	55.4	69.3	71.1	54.5	73.6	61.3
P_3	93.3	39.2	68.4	60.2	100.0	34.0	57.3	71.4	85.2	58.3	74.9	63.4
P_4	89.6	34.4	64.0	59.9	100.0	35.1	65.8	71.0	77.8	72.5	75.6	55.8
P_5	91.1	60.9	56.9	64.2	99.3	45.7	77.9	77.1	67.4	79.9	75.1	58.3
P_6	94.1	66.9	72.1	55.5	100.0	53.0	77.2	74.3	84.4	74.0	74.5	59.2
P_7	94.8	73.6	85.4	53.6	98.6	61.6	80.4	81.5	92.6	86.5	76.0	62.0
P_8	94.8	74.6	85.2	85.0	98.6	69.1	79.3	75.2	87.4	92.9	74.2	54.5
P_9	91.1	76.9	91.3	89.4	97.9	65.4	77.6	83.9	93.3	85.8	76.1	58.9
μ	92.7	54.8	70.8	65.6	99.4	44.1	66.7	71.8	77.6	71.7	75.00	59.60
δ	1.73	17.34	12.22	12.09	0.77	17.25	14.61	8.97	13.36	14.75	0.91	2.83

- Average Forgetting Rate Fr_{avg} is used to evaluate the actual difficulty of each task based on the final result.

4.5 The effectiveness of KB-C

Prior Difficulty: $D^{j}_{prior} := \frac{1}{K-1}\sum_{j=1;j\neq i}^{K} S^{j}_{i}$

Posterior Difficulty: $D^{j}_{post} := Fr^{i}_{avg}$;

		T_0	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	PCC_s
D_{prior}		0.121	0.141	0.168	0.054	0.035	0.186	0.112	0.152	0.146	0.137	-
D_{post}	EA-EMR	0.060	0.098	0.061	0.028	0.001	0.137	0.139	0.060	0.044	0.051	0.559
	MLLRE	0.022	0.078	0.091	0.085	-0.004	0.147	0.060	0.064	0.069	0.075	0.667
	EMAR	0.036	0.016	0.027	0.007	0.006	0.016	0.008	0.026	0.020	0.005	0.499
	CML	-0.002	0.108	0.051	0.065	-0.002	0.113	0.108	0.046	0.070	0.068	**0.470**

Summary: domain/range-based relation similarity is positively related to the difficulty of the task； Our Kb-C module reduces the interference of similar tasks.

The Next

Part 5

Conclusion

8 Conclusion

A Task: Continual Relation Extraction;

Two Problems: Catastrophic Forgetting & Order-Sensitivity;

Two Factors: Over-Fitting & Task Similarity;

Two Main Contribution: CML Framework and KB-C Module;

Three Future Works.

References

[1] Yu et al.: Improved neural relation detection for knowledge base question answering. Proceedings of ACL. pp. 571–581 (2017)
[2] Wang et al.: Sentence embedding alignment for lifelong relation extraction. Proceedings of NAACL-HLT2019. pp. 796–806 (2019)
[3] Obamuyide et al.: Meta-learning improves lifelong relation extraction. Proceedings of RepL4NLP@ACL2019. pp. 224–229 (2019)
[4] Han et al.: Continual relation learning via episodic memory activation and reconsolidation . Proceedings of ACL2020. pp. 6429–6440 (2020)
[5] Yoon et al.: Scalable and order-robust continual learning with additive parameter decomposition. arXiv: Learning (2020)
[6] Chen, Z., Liu, B.: Lifelong Machine Learning. Morgan & Claypool Publishers (2016)
[7] Lopez-Paz et al.: Gradient episodic memory for continual learning. Proceedings of NeurIPS2017. pp. 6467–6476 (2017)
[8] Kirkpatrick et al.: Overcoming catastrophic forgetting in neural networks. CoRR abs/1612.00796 (2016)
[9] Liu et al.: Rotate your networks: Better weight consolidation and less catastrophic forgetting. Proceedings of ICPR. pp. 2262–2268 (2018)

THANKS

2020.12.19