Classifying Math Knowledge Components via Task-Adaptive Pre-Trained BERT

Jia Tracy Shen, Michiharu Yamashita, Ethan Prihar, Neil Heffernan, Xintao Wu, Sean Mcrew, Dongwon Lee

June 14-18, 2021

Introduction

- Motivation
- Our Approach
- Evaluation
- Conclusion

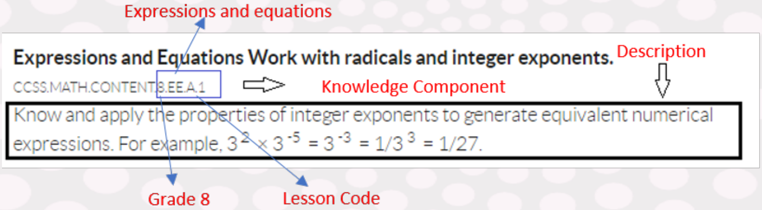

Figure – An Example of Knowledge Component and Its description

Motivation

1. Identifying Knowledge Component (KC) is tedious &challenging to ITS, LMS, Teachers (see in below figure)
2. Limits of Prior work : small scale KCs, use single type data, not using NLP approach
3. Predict 3 tasks based on description, video title and problem texts

Figure – Three tasks to identify KC

Task-adpative Pre-trained BERT (TAPT) : pre-train on task-specific data

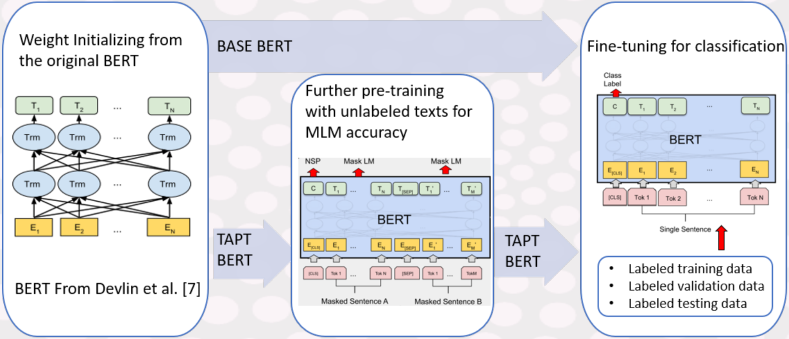

Figure – An illustration of training and fine-tuning process of BASE vs. TAPT

Our Approach -ii

Table – A summary statistics of datasets.

Name	# Labels	# Texts	# Tokens	Fine-tuning Partition		
				Training (72%)	Validation (8%)	Testing (20%)
D_d	385	6,384	84,017	4,596	511	1,277
D_t	272	6,748	62,135	4,858	540	1,350
D_p	213	13,722	589,549	9,879	1,098	2,745
D_{d+t}	/	13,132	146,152	/	/	/
D_{d+p}	/	20,106	673,566	/	/	/
D_{t+p}	/	20,470	651,684	/	/	/
D_{all}	/	26,854	735,701	/	/	/

Our Approach -iii

TAPT outperforms 6 baselines as shown below :

Table – Accuracy comparison ($BL\dagger$ for baseline best, and * for statistical significance with p-value < 0.001)

Approach Type	Algorithm	D_d		D_t		D_p	
		Acu@1	Acu@3	Acu@1	Acu@3	Acu@1	Acu@3
Classical ML	SVM [Karlovčec et al., 2012]	44.87	70.40	48.15	70.30	78.07	87.69
	XGBoost	43.07	71.34	45.33	66.15	77.63	87.94
	Random Forest	49.26	78.78	49.33	74.37	78.03	88.23
Prior Work	Skip-Gram NN [Pardos, 2017]	34.07	34.15	43.00	43.52	76.88	77.06
	Sklearn *MLP* [Patikorn et al., 2019]	50.53	74.41	48.22	57.95	80.70	81.13
BERT	BASE	48.30	76.40	50.99	76.55	81.73	90.99
	TAPT	**50.60**	**79.29**	**52.71**	**78.83**	**82.43**	**92.51**
Improvement	$\lvert TAPT - BL\dagger \rvert$	0.07	0.51	1.72	2.28	0.70	1.52
	$\lvert TAPT - BASE \rvert$	2.30*	0.51*	1.72*	2.28*	0.70*	1.52*

Our Approach -iv

Pre-train with augmented data :

Table – Acu@3 : BASE vs. TAPT. (best and 2nd best per row in bold and underlined, and subscripts indicate outperformance over BASE)

Data	BASE	Simple			Augmented			
		$TAPT_d$	$TAPT_t$	$TAPT_p$	$TAPT_{d+t}$	$TAPT_{d+p}$	$TAPT_{t+p}$	$TAPT_{all}$
D_d	76.40	$79.29_{2.89}$	$78.78_{2.38}$	$77.84_{1.44}$	$\underline{79.40}_{3.00}$	$\mathbf{79.56}_{3.16}$	$79.01_{2.61}$	$79.01_{2.61}$
D_t	76.55	$\underline{77.85}_{1.30}$	$\mathbf{78.83}_{2.28}$	$76.30_{-0.25}$	$77.56_{1.01}$	$77.56_{1.01}$	$77.70_{1.15}$	$77.78_{1.23}$
D_p	90.99	$91.22_{0.23}$	$91.44_{0.45}$	$\underline{92.51}_{1.52}$	$92.06_{1.07}$	$92.50_{1.51}$	$\mathbf{92.64}_{1.65}$	$92.35_{1.36}$

Evaluation-TEXSTR

$TEXSTR : \Lambda = \alpha \cdot C_t + (1 - \alpha) \cdot C_s$, where α controls the weight between C_t (semantic sim.) and C_s (structural sim.) as an oscillating parameter. A threshold value {0.5, 0.75, 0.9} is applied on the Λ results to set the criteria on reconsider miss-predictions as correct.

Table – % of miss-predictions recovered by $TEXSTR$ (Λ)

Data	# Miss-predictions	$\Lambda > 0.5$			$\Lambda > 0.75$			$\Lambda > 0.9$		
		$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$
D_d	248	70.16	68.95	72.98	52.82	**24.19**	8.87	32.26	**2.42**	0.81
D_t	240	58.33	55.83	57.5	37.92	**17.08**	6.67	17.08	**0**	1.25
D_p	166	60.84	56.63	58.43	38.55	**16.27**	5.42	18.67	**1.2**	1.2

Table – % of top-3 predictions by relevance (Υ) level when $\alpha = 0.5$

Υ	Top 1			Top 2			Top 3		
	Λ	Teachers	Δ	Λ	Teachers	Δ	Λ	Teachers	Δ
> 0.5	100	54.31	**-45.69**	100	40.95	**-59.05**	100	21.98	**-78.02**
> 0.75	37.93	43.53	**+5.60**	20.69	27.16	**+6.47**	6.9	13.79	**+6.89**
> 0.9	3.45	31.03	**+27.58**	0	13.79	**+13.79**	0	9.48	**+9.48**

Conclusion & Thank you

1. TAPT is the first NLP model classifying full set KC (385) and achieved a new record by outperforming six baselines by up to 2% at Acu@1 and up to 2.3% at Acu@3.
2. TAPT trained on the augmented data by combining different task-specific texts had better Acu@3 than TAPT simply trained on the individual datasets.
3. Our new evaluation measure TEXSTR was able to reconsider 56-73% of miss-predictions as correct for practical use.
4. Source code and slide : https ://github.com/tbs17/TAPT-BERT/tree/master
5. Find more about our research at Pike Group @Penn State : http ://pike.psu.edu/

References

[Karlovčec et al., 2012] Karlovčec, M., Córdova-Sánchez, M., and Pardos, Z. A. (2012).
Knowledge component suggestion for untagged content in an intelligent tutoring system.
Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7315 LNCS :195–200.

[Pardos, 2017] Pardos, Z. A. (2017).
Imputing KCs with Representations of Problem Content and Context.
In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 148–155.

[Patikorn et al., 2019] Patikorn, T., Deisadze, D., Grande, L., Yu, Z., and Heffernan, N. (2019).
Generalizability of methods for imputing mathematical skills needed to solve problems from texts.
Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11625 LNAI :396–405.