Learning Outcomes and Their Relatedness Under Curriculum Drift

Sneha Mondal^{1(✉)}, Tejas I. Dhamecha^{1(✉)}, Smriti Pathak², Red Mendoza³,
Gayathri K. Wijayarathna³, Paul Gagnon³, and Jan Carlstedt-Duke³

¹ IBM Research, Bangalore, India

{snemonda, tidhamecha}@in.ibm.com

² Imperial College, London, UK

³ Lee Kong Chian School of Medicine,
Nanyang Technological University, Singapore, Singapore

Abstract. A typical medical curriculum is organized as a hierarchy of learning outcomes (LOs), each LO is a short text that describes a medical concept. Machine learning models have been applied to predict relatedness between LOs. These models are trained on examples of LO-relationships annotated by experts. However, medical curricula are periodically reviewed and revised, resulting in changes to the structure and content of LOs. This work addresses the problem of model adaptation under curriculum drift. First, we propose heuristics to generate reliable annotations for the revised curriculum, thus eliminating dependence on expert annotations. Second, starting with a model pre-trained on the old curriculum, we inject a task-specific transformation layer to capture nuances of the revised curriculum. Our approach makes significant progress towards reaching human-level performance.

Keywords: Learning outcome(s) · Data drift · Model adaptation

1 Introduction

The LO-relationship extraction task, recently introduced in [8], seeks to predict the degree of relatedness between learning outcomes (LOs) in a curriculum. The authors examine the curriculum of the Lee Kong Chian School of Medicine, which spans five years of education and covers about 4000 LOs; each LO is a short statement describing a concept that students are expected to master. A hierarchy, designed by curriculum experts, groups these LOs at different levels of granularity. A successful clinical encounter requires students to conceptually relate and marshal knowledge gained from several LOs, spread across years and across distant parts of the curriculum hierarchy. This underscores the need for an automatic LO-relationship extraction tool (hereafter called LReT).

In our earlier work [8], this is abstracted as a classification task, where a pair of LOs is categorized as being strongly related (high degree of conceptual

S. Mondal and T. I. Dhamecha—Contributed equally.

© Springer Nature Switzerland AG 2020

I. I. Bittencourt et al. (Eds.): AIED 2020, LNAI 12164, pp. 214–219, 2020.

https://doi.org/10.1007/978-3-030-52240-7_39

similarity), weakly related (intermediate conceptual similarity), or unrelated (no conceptual similarity). An LReT is trained on annotated data obtained from subject matter experts (SMEs), who are both faculty and doctors.

However, this curriculum is periodically reviewed and revised. Modifications are made to both content (emphasising some LOs, dropping others, merging a few), as well as organization (grouping LOs differently, re-evaluating classroom hours dedicated to each). Table 1 compares an old LO with its revised counterpart. Note that the textual formulation (hence underlying concept) of the LO has been modified. Additionally, the LO has been re-grouped under a separate set of verticals - **Longitudinal Course**, **Module**, and **Assessment Type**, while doing away with **Clinical Block**, the only vertical in the previous version.

Table 1. Semantic and curricular change in LOs after curriculum revision

	Old curriculum	Revised curriculum
Learning Outcome (LO)	Identify the main mineralocorticoid in humans and describe its principal actions	Describe the synthesis, principal actions and control mechanisms of glucocorticoids, mineralocorticoids and adrenal steroids
Theme	Scientific Basis of Medicine	Scientific Basis of Medicine
Fundamental	Structure & function of the human body in health (HSF)	Structure & function of the human body in health (HSF)
Fundamental Unit	Anatomy, Physiology, Biochemistry	Biochemistry
Clinical Block	Renal and Endocrine	–
Longitudinal Course	–	Foundations of Clinical Practice (FCP)
Module	–	Endocrine System
Assessment Type	–	Written Exercise, Oral Presentation

As the curriculum drifts, so do relationships between its constituent LOs. An LReT trained on one version of the curriculum may not perform well on the revised version. Re-obtaining SME annotations carries appreciable cognitive and cost overheads, making it impractical to train an LReT from scratch.

We present a systematic approach towards LO-relationship extraction under curriculum drift. Beginning with the SME-labelled dataset on the old curriculum, we employ heuristics to create a pseudo-labelled dataset for the revised curriculum. With some supervision now available, we tune the existing pre-trained model to the nuances of the revised curriculum, and compare its efficacy against human performance.

This aligns with existing work on domain adaptation and transfer learning [6, 10]; both study scenarios where training and test data do not derive from the same distribution. In contrast, not only do we adapt the model to a modified domain, but also generate data pertinent to this domain, thus eliminating the need for human intervention. This bridges the gap between building a reliable LReT, and deploying it against a changing curriculum landscape.

2 Silver Standard Dataset Generation

Starting with SME-annotated old LO pairs, which serves as the gold-standard dataset, we proceed in two steps. First, we define a *mapping* that links an LO from the old curriculum (OC) to its closest matching counterpart in the revised curriculum (RC):

$$M(p) = \{r | \text{sim}(p, r) \geq \text{sim}(p, r'), \forall r' \in RC\}, p \in OC. \quad (1)$$

where *sim* is an appropriate semantic textual similarity metric. Intuitively, the mapping score, $\text{sim}(p, M(p))$ captures the extent of semantic drift in the content of an LO.

Fig. 1. (A) **base-model** trained on gold-standard data from OC. (B) Model trained from **Scratch** on silver-standard data from RC. (C) Manually map features (**MF**) from RC to OC, and then use **base-model**. (D) Learn a feature transform (**FT**) from RC that approximates OC-like features by leveraging weak correspondence between RC and OC. The base-model can be further smoothed (**FT-S**).

Thereafter, we rely on *pruning*. Recall that the gold-standard dataset ($\mathcal{D}_{old}$) consists of old LO pairs (p, q), along with an SME-annotated class label. A silver-standard dataset for the revised curriculum ($\mathcal{D}_{rev}$) is derived by pruning the mapping scores of an old LO pair at a pre-defined threshold (τ), while retaining its class label. Formally,

$$\mathcal{D}_{rev}(\tau) = \{(M(p), M(q), \text{label}) \mid \text{sim}(p, M(p)) \geq \tau \text{ and } \text{sim}(q, M(q)) \geq \tau\} \\ \forall (p, q, \text{label}) \in \mathcal{D}_{old}. \quad (2)$$

Effectively, we propagate the SME-label from a LO pair in old curriculum to their corresponding maps in the revised curriculum, only if the both mapping scores exceed the threshold. These pseudo-labeled instances constitute the silver-standard dataset.

3 Proposed Model Adaptation Approaches

The **base-model** (Fig. 1(A)), trained on gold-standard LO pairs of the old curriculum, predicts posterior probabilities for **Strong**, **Weak**, and **None** classes. As a comparative baseline, we train a model from scratch on the silver-standard dataset, without leveraging the **base-model**. We then explore three approaches to adapt **base-model**:

1. **Manual Feature Mapping (MF)**, where we manually map features from the revised curriculum to the old curriculum, and drop features that cannot be mapped (Fig. 1(C)). The resultant feature set can be fed to the **base-model** for predicting LO relatedness in new curriculum.
2. **Feature Transformation (FT)**: In this novel approach (Fig. 1(D)), we inject a fully connected layer that transforms the revised feature set to an *approximate* old feature set, which can then be fed to the **base-model**. The silver standard dataset is leveraged to train only this transformation layer, i.e. **base-model** layers are frozen.
3. **Feature Transformation with Smoothing (FT-S)**: Once the transformation weights converge to an extent, we unfreeze the **base-model** parameters and train for a few epochs to allow fine-grained updates to the entire network.

4 Experiments and Analysis

Table 2a compares model adaptation techniques outlined in Sect. 3. All approaches that leverage the **base-model** outperform training from **Scratch**, to various degrees. Feature transformation with smoothing (**FT-S**) yields the highest macro-F1, thus establishing that a) the **base-model** encodes some task-specific information independent of the specific curriculum, b) the revised feature-set can be adequately modeled as a linear transformation of the old feature-set, and c) additional smoothing over parameters of the **base-model** allows it to learn curriculum-specific nuances.

Furthermore, as shown in Table 2b, the high variance in model performance stems from the small size of training and test sets for each cross-validation split, and the macro-F1 score is sensitive to samples in the specific test split. We perform paired t-test to ascertain that except for two pairs, FT vs MF ($p = 6.8 \times 10^{-2}$) and FT vs FT-S ($p = 6.6 \times 10^{-2}$), differences between all other technique-pairs are statistically significant at 95% confidence interval.

Finally, for a small held-out set ($n = 229$), we obtain annotations separately from two SMEs and compute the inter-annotator agreement (71.7% macro-F1), which serves as a skyline. As shown in Table 2d, considering one SME as ground-truth and comparing against FT-S’s predictions, the human-machine agreement turns out to be 64.4%. Compared to human performance, our reported results are moderately high, with, of course, some further scope of improvement.

Table 2. (a) Macro-average F1 (mean $\pm$ standard deviation in %) for **base-model** and several model adaptation approaches (b) per fold Macro-F1 (in %) of Scratch vs FT-S on SS-5. Note the high variance across splits, but consistent improvement of FT-S over Scratch. (c) Human vs human and (d) human vs algorithm agreement.

(a)						(b)		
Dataset (τ)	Size	Scratch	MF	FT	FT-S	Fold	Scratch	FT-S
SS-1(0.85)	195	35.5 $\pm$ 10.8	45.3 $\pm$ 10.8	54.0 $\pm$ 16.5	54.4 $\pm$ 12.5	1	25.7	49.4
SS-2(0.80)	315	35.0 $\pm$ 10.3	46.2 $\pm$ 10.7	50.7 $\pm$ 12.8	54.0 $\pm$ 12.3	2	23.9	48.3
SS-3(0.75)	466	35.1 $\pm$ 7.6	46.7 $\pm$ 9.1	50.1 $\pm$ 8.2	51.4 $\pm$ 6.8	3	29.2	53.9
SS-4(0.70)	692	35.7 $\pm$ 7.7	46.0 $\pm$ 7.0	49.4 $\pm$ 9.6	49.4 $\pm$ 8.8	4	48.9	54.7
SS-5(0.65)	963	35.6$\pm$ 8.6	48.1 $\pm$ 4.9	50.4 $\pm$ 5.1	52.0$\pm$ 4.2	5	42.1	53.7
Macro-average F1 of base-model on gold standard : 59.2 $\pm$ 1.2						6	44.3	55.3
Gold-standard : 2,276 LO pairs (Strong : 314, Weak : 471, None : 1,491)						7	32.7	48.8
						8	43.4	57.1
						9	35.6	55.4
						10	29.7	43.7
						-	35.6$\pm$8.6	52.0$\pm$4.2

(c)					(d)					
Confusion Matrix		Annotator 2				Confusion Matrix		FT-S		
		Strong	Weak	None				Strong	Weak	None
Annotator 1	Strong	30	10	0	40	Annotator 1	Strong	29	3	8
	Weak	7	24	32	63		Weak	15	17	31
	None	0	1	125	126		None	2	1	123
Total		37	35	157	229	Total		46	21	162
Macro-Average F1= 71.74, Accuracy= 78.16					Macro-Average F1= 64.4, Accuracy= 73.7					

References

1. Bjerva, J., Kouw, W., Augenstein, I.: Back to the future-sequential alignment of text representations. arXiv preprint [arXiv:1909.03464](https://arxiv.org/abs/1909.03464) (2019)
2. Chan, J., Bailey, J., Leckie, C.: Discovering correlated spatio-temporal changes in evolving graphs. Knowl. Inf. Syst. **16**(1), 53–96 (2008)
3. Chen, Y., Wuillemin, P.H., Labat, J.M.: Discovering prerequisite structure of skills through probabilistic association rules mining. International Educational Data Mining Society (2015)
4. Gravemeijer, K., Rampal, A.: Mathematics curriculum development. In: Cho, S.J. (ed.) The Proceedings of the 12th International Congress on Mathematical Education, pp. 549–555. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-12688-3_57
5. Käser, T., Klingler, S., Schwing, A.G., Gross, M.: Beyond knowledge tracing: modeling skill topologies with bayesian networks. In: Trausan-Matu, S., Boyer, K.E., Crosby, M., Panourgia, K. (eds.) ITS 2014. LNCS, vol. 8474, pp. 188–198. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07221-0_23
6. Kouw, W.M., Loog, M.: A review of domain adaptation without target labels. IEEE Trans. Pattern Anal. Mach. Intell. (2019)
7. Kumar, I., Balakrishnan, S.: Beyond basic: a temporal study of curriculum changes in a first-year communication course. Int. J. Res. Bus. Stud. **4**, 14 (2019). ISSN 2455–2992
8. Mondal, S., et al.: Learning outcomes and their relatedness in a medical curriculum. In: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 402–411 (2019)

9. Pyysalo, S., Ginter, F., Moen, H., Salakoski, T., Ananiadou, S.: Distributional semantics resources for biomedical text processing (2013)
10. Raina, R., Battle, A., Lee, H., Packer, B., Ng, A.Y.: Self-taught learning: transfer learning from unlabeled data. In: Proceedings of the 24th International Conference on Machine Learning, pp. 759–766 (2007)
11. Reis, S.: Curriculum reform: why? what? how? and how will we know it works? *Isr. J. Health Policy Res.* **7**, 30 (2018). <https://doi.org/10.1186/s13584-018-0221-4>
12. Stankov, S., Rosić, M., Žitko, B., Grubišić, A.: Tex-sys model for building intelligent tutoring systems. *Comput. Educ.* **51**(3), 1017–1036 (2008)
13. Zouaq, A., Nkambou, R.: Building domain ontologies from text for educational purposes. *IEEE Trans. Learn. Technol.* **1**(1), 49–62 (2008)