Learning Outcomes and Their Relatedness in a Medical Curriculum

Sneha Mondal[1], **Tejas Indulal Dhamecha**[1], **Shantanu Godbole**[1], **Smriti Pathak**[2], **Red Mendoza**[3], **K Gayathri Wijayarathna**[3], **Nabil Zary**[3], **Swarnadeep Saha**[1], **Malolan Chetlur**[1]
[1]IBM Research - India
[2]Imperial College, London
[3]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore
{snmondal, tidhamecha, shantanugodbole, swarnads, mchetlur}@in.ibm.com,
smriti.pathak@kcl.ac.uk, {mendozaredante, gwkumari, nabil.zary}@ntu.edu.sg

Abstract

A typical medical curriculum is organized in a hierarchy of instructional objectives called Learning Outcomes (LOs); a few thousand LOs span five years of study. Gaining a thorough understanding of the curriculum requires learners to recognize and apply related LOs across years, and across different parts of the curriculum. However, given the large scope of the curriculum, manually labeling related LOs is tedious, and almost impossible to scale. In this paper, we build a system that learns relationships between LOs, and we achieve up to human-level performance in the LO relationship extraction task. We then present an application where the proposed system is employed to build a map of related LOs and Learning Resources (LRs) pertaining to a virtual patient case. We believe that our system enables building educational tools to help medical students grasp the curriculum better, within classroom and Intelligent Tutoring Systems (ITS) settings.

1 Introduction

Learning Outcomes (LOs) encapsulate discrete knowledge components and provide a framework for curriculum planning, teaching, learning, and assessment. In this work, we study the curriculum of the Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore. At the highest level, their curriculum is organized into major *Themes*, which branch into *Fundamentals*, and further into *Fundamental Units*. A Fundamental Unit is comprised of multiple related *Topics*, and each topic constitutes several LOs. Thus, related LOs get grouped together at multiple levels of increasing granularity. This hierarchy is hand-curated by medical experts and represents a well-formed, well-understood body of knowledge.

However, qualitative evidence suggests that significant relationships exist between LOs placed far apart in the curriculum; these relationships cannot be uncovered without explicit expert intervention. Figure 1 illustrates one such instance, where LOs drawn from disjoint sections of the curriculum hierarchy are related as they address multiple aspects of HIV treatment.

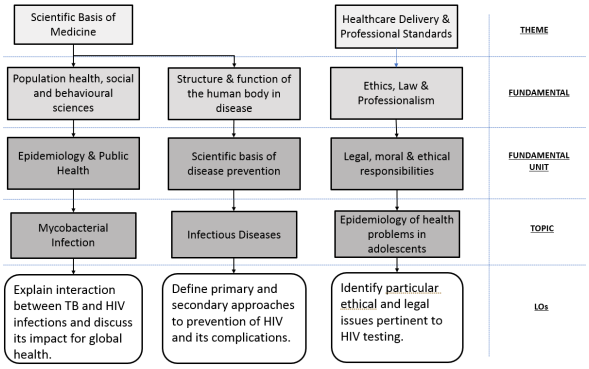

Figure 1: Related Learning Outcomes placed far apart in the expert-curated curriculum hierarchy.

Our main motivation in this work is to automatically discover LO relationships that cannot be accessed by a straightforward navigation of the curriculum. Extracting such LO relationships can help build a knowledge-base that can be foundational to various educational tools. To this end, we propose looking into the semantic content of disparate LOs, in addition to their relatedness specified by the curriculum hierarchy. We formulate this as a three-class classification task. Given a pair of LOs, they are categorized as being either strongly related or weakly related, or unrelated.

Although the current study is limited to a medical curriculum, our approach is general. Techniques reported in this paper would extend to any curricula that take a 'design down' approach (Harden, 2002), where related LOs are nested in a hierarchical order. An LO-relationship extraction tool that utilizes both semantic and curricu-

Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 402–411
Florence, Italy, August 2, 2019.

lum cues, can be exploited by Intelligent Tutoring Systems (ITS) to suggest useful interventions to both learners and instructors. Potential applications include: a) improved content recommendation, by proactively suggesting pre-requisites or guiding the learner to discover LOs that are related across disjoint sections of the curriculum; b) designing better assessment items which test a learner on closely related LOs; and c) accurate learner modeling, by taking into account all related LOs when tracking the progress of a learner's mastery of an LO. Building upon these motivations, this work documents our efforts to answer the following research questions :

RQ1: *Which features determine relatedness between LOs?* Information available to us is both structured (by way of a well-defined curriculum hierarchy), as well as unstructured (by way of free text descriptions of LOs). We aim to devise a method to appropriately integrate the two in order to compare two LOs.

RQ2: By design, LOs are crisp and compact. A drawback of their conciseness is that they do not provide enough information to ascertain relatedness with other, similarly concise LOs. So, we ask, *can the resources linked to LOs be suitably leveraged to improve the quality of LO-relationship extraction?*

RQ3: *Are there any latent factors beyond curriculum and semantic similarity establishing relatedness among LOs?* If so, are they exploited by the proposed approach?

RQ4: *Can LO relatedness be used to understand a virtual patient case?* Disparate LOs of disease and symptoms could be related in the context of a case. We leverage the LO relationship extraction system to understand the context of a case, and build a case map from relevant concepts.

2 Related Work

Intelligent Tutoring Systems (ITS) greatly improve students' user experience, even in comparison to human tutors (Aleven et al., 2004; VanLehn, 2011). Automated methods for creating domain ontology from text have been explored in (Zouaq and Nkambou, 2008). While most previous work employ semantic networks with frames and production rules (Stankov et al., 2008), we tap into state-of-the-art AI - based techniques to *learn* semantic relationship between LOs, as opposed to enumerating rules to generate them. Our work comes close in spirit to that of (John et al., 2015), that seek to generate knowledge graphs for closely related math word problems. They employ a random-walk paradigm on a graph whose edges are weighed by tf-idf based cosine similarity. Unlike them, we exploit the existing medical curriculum hierarchy, and use a suite of semantic features extending beyond tf-idf.

Graphs have been widely used to establish prerequisite relationships between domain knowledge concepts (Chen et al., 2015; Käser et al., 2014), where a link between concepts indicates a prerequisite-outcome relation. Guerra et al. (2015) represent a student model as a graph where links are gradually added between pairs of knowledge concepts when a student is able to work with aforementioned pairs in the same context. Similarly, Rihák and Pelánek (2017) group similar knowledge concepts using learners' performance data and response time metadata. However, missing here are relations between knowledge concepts already encoded in the curriculum and its textual content.

There is a parallel thread of work on Semantic Textual Similarity (STS), which measures the degree of equivalence in the underlying semantics of paired snippets of text (Agirre et al., 2015, 2016, 2012). This aligns with our work since it is also posed as a natural language understanding problem. However, techniques explored within the ambit of STS are agnostic to any domain specific ontology. This is a major drawback for our application, as the medical curriculum embodies pertinent domain information, which, as we later show, goes a long way in establishing accurate relationships between LOs.

To the best of our knowledge, we are the first to exploit expert-annotated data from an extensively detailed medical curriculum for the LO-relationship extraction task. By establishing semantic relationships among the curriculum concepts, we bridge the gulf between hand-curated domain-specific ontologies and state-of-the-art data driven textual similarity measures, and show its utility in understanding a patient case.

3 Curriculum and Problem Statement

In this section, we briefly describe the organization of LOs in the medical curriculum, and formulate the problem statement.

Medical Curriculum : The curriculum content is designed around 3 *Themes* that run throughout the programme: 1) Scientific Basis of Medicine,

Relation with Ref. LO	LO	Fundamental unit	Fundamental
-	**Reference LO:** Explain the normal development of the embryonic heart	Embryology	Human Structure & Function
`Strong`	**LO1:** Explain how the pulmonary and systemic circulations are linked in fetal life	Embryology	Human Structure & Function
`Weak`	**LO2:** Explain the mechanisms underlying Starling's Law of the Heart	Anatomy/ Physiology	Human Structure & Function
`None`	**LO3:** List the clinical uses of pulse oximetry	History, Exam. and MSE	Integrated Clinical Practice

Table 1: Example LO relationships along with their placement in curriculum hierarchy.

2) Clinical Management and Patient Centred Care, and 3) Healthcare Delivery and Professional Standards. The themes correspond to cognition, attitude, and skills of the spiral curriculum, as suggested by Harden (1999).

Figure 1 depicts the organization of LOs into themes that consist of *Fundamentals*, branching in order into *Fundamental Units* and *Topics*. Additionally, an LO is not constrained to belong to a unique fundamental unit, and may span a small *set* of relevant themes, fundamentals and fundamental units. Overall, our curriculum contains $4,251$ LOs, organized into 670 Topics, 81 Fundamental Units, and 16 Fundamentals.

Learning Outcomes and Resoures : In addition, curriculum designers have manually linked a majority of the LOs to relevant study material, termed Learning Resource (LR). These LRs could be selected pages from textbooks, transcripts of video lectures, links to online reading material, or extracts from presentations. In this work, we restrict ourselves to LRs that are well-curated slide decks, in the form of pdf files. We also note that all LOs and corresponding LRs are authored in English.

3.1 Problem Statement

Since our goal is to predict the degree of relatedness between a pair of LOs, we define our problem statement as follows: Given two LOs and their positions in the curriculum, classify the relationship between them as `Strong`, `Weak`, or `None`. More precisely, we seek to learn a function that, for a pair of LOs p and q, maps them to one of three possible classes, i.e.,

$$f : (p, q) \rightarrow \{\texttt{Strong}, \texttt{Weak}, \texttt{None}\}$$

Such a function could then be employed to predict relationships between any unseen pair of LOs.

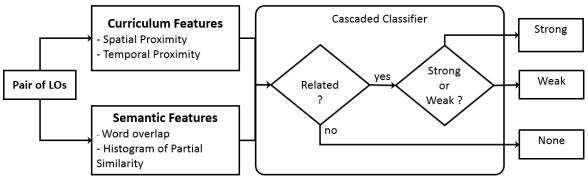

Figure 2: Proposed approach to classify an LO pair.

Expert-annotated Data : Annotations are obtained from Subject Matter Experts (SMEs), who are both doctors and faculty. The annotated data consists of pairs of LOs, each pair labeled as `Strong`, `Weak`, or `None`. Additionally, SMEs were requested to provide guided comments to help us understand their mechanism for coping with ambiguity. For a subset of LO-pairs, annotations were obtained separately from two SMEs to ascertain inter-annotator agreement.

4 Proposed Solution

Our approach for LO-relationship extraction is summarized in the block diagram in Figure 2. The pipeline involves choosing features meaningful for the task, followed by a cascaded classifier design. In sections that follow, we detail and motivate individual components of the pipeline. Subsequently, we investigate the benefit of employing LRs and additional metadata in our task.

4.1 Features

Observations from data indicate that two classes of features, curriculum and semantic, are critical for LO-relationship extraction. Thus, final representation for a pair of LOs is the concatenation of their curriculum and semantic features.

4.1.1 Curriculum-based Features

Curriculum-based features characterize the relative position of LOs within the curriculum hierarchy, which is used to obtain spatial and temporal proximity estimates, as follows:

Spatial Proximity: We hypothesize that the closer LOs are located in the curriculum hierarchy, the more likely they are to be related. In Table 1, we compare a reference LO against LOs that are placed gradually farther in the curriculum (same fundamental unit, separate fundamental units within the same fundamental, and separate fundamentals, respectively). In this specific example, we note that the degree of relatedness decreases with decreasing spatial proximity from the reference LO.

As discussed in previously, each LO may span multiple fundamentals, fundamental units, and themes. For LOs p and q, let their set of themes, fundamentals, and fundamental units be $T_{i\in\{p,q\}}$, $F_{i\in\{p,q\}}$, and $U_{i\in\{p,q\}}$, respectively. Proximity of the LO pair is represented as:

$$SP(p,q) = [J(T_p,T_q), J(F_p,F_q), J(U_p,U_q)]$$

where Jaccard similarity J between sets A and B is defined as $J(A,B) = |A \cap B|/|A \cup B|$.

Temporal Proximity: Related concepts are taught successively within a course curriculum, hence, the time of delivery of LOs is an indicator of their relatedness. For LOs p and q, let $y_{i\in\{p,q\}}$ and $w_{i\in\{p,q\}}$ be their year and week of delivery, respectively. Temporal proximity is then encoded as

$$TP(p,q) = [|y_p - y_q|, |w_p - w_q|]$$

The year information is encoded separately since curriculum focus differs year-wise, i.e., content taught in the last week of year 1 may not always be related to first week of year 2.

4.1.2 Semantic Features

While curriculum hierarchy encodes one paradigm for grouping related LOs, it misses the rich semantic information contained in the text of the LO. Revisiting our example, consider the LOs : 1) *Identify the particular ethical and legal issues pertinent to HIV testing.* and 2) *List some of the common HIV indicator conditions and HIV-related opportunistic infections*. They are far apart in the hierarchy, however there exists a `Strong` relationship between them as they are related in the context of treating an HIV-infected patient. Thus we explore features that encode semantic similarity between LOs.

Embedding based Features: Semantic relatedness between LOs is often encapsulated by the similarity of their constituent tokens. As an example, the LOs : 1) *List the common symptoms of sudden cardiac arrest*, and 2) *List the common symptoms of myocardial infarction*, are related since term pairs (*cardiac*, *myocardial*) and (*arrest*, *infarction*) refer to similar entities. Since exact token matching (as in Eq. 2) is deficient in modelling such semantic overlap, we utilize word embeddings (Chiu et al., 2016) to represent individual tokens, which are further used to compute the following similarity measures.

• **Word Overlap:** Each LO text is treated as a bag-of-words. We define that a word w_i in LO p overlaps with a word w_j in LO q, if their cosine similarity in the word embedding space exceeds a certain threshold δ. Based on this *soft* matching of words, we define semantic word overlap to be the fraction of matching word pairs across the two bags-of-words, as:

$$WO(p,q) = \frac{\sum_{w_i\in p}\sum_{w_j\in q} \mathbb{1}\left[\cos(\mathbf{w_i},\mathbf{w_j}) \geq \delta\right]}{|p||q|} \quad (1)$$

where $\mathbb{1}[\cdot]$ is an operator that evaluates to 1 if corresponding condition is True, and 0 otherwise.

• **Histogram of Partial Similarities (HoPS):** We employ HoPS (Saha et al., 2018) to model the *similarity profile* between two LOs. For each word w_i in LO p, first its similarity score is computed with respect to LO q as:

$$S(w_i,q) = \max_{w_j\in q} \cos(\mathbf{w_i},\mathbf{w_j}), \text{ where } w_i \in p$$

This strategy pairs each word in LO p with its closest matching counterpart in LO q. These similarity scores are then partitioned into N bins and normalized, resulting in a histogram of scores for p. We obtain another normalized histogram by binning the similarity scores for each word in q with respect to p.

$$HoPS(p,q) = [\text{Histogram}(\{S(w_i,q)|w_i \in p\}) \quad , \\ \text{Histogram}(\{S(w_j,p)|w_j \in q\})]$$

Unlike word overlap, HoPS considers all tokens in the LO text without thresholding on a similarity score, and hence provides a more granular similarity profile between LOs.

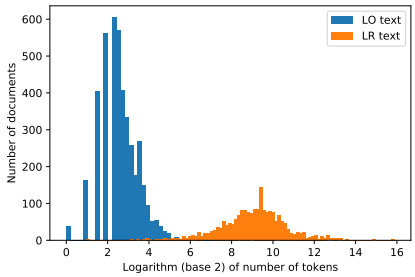

Figure 3: Distribution of the length of LO and LR texts on a logarithmic scale.

TF-IDF based Feature: For LOs p and q, let $\mathbf{p} \in R^{|v|}$ and $\mathbf{q} \in R^{|v|}$ be their respective representations as tf-idf vectors (Ramos et al., 2003); where v is the vocabulary set. Then tf-idf based similarity is encoded as their absolute difference and Hadamard product

$$TI(p, q) = [|\mathbf{p} - \mathbf{q}|, \mathbf{p} \circ \mathbf{q}] \quad (2)$$

Essentially, our representation encodes the information gap between LOs in a pair, in terms of their exact token overlap weighed by importance of said token in the LO corpus.

The final feature representation for an LO pair is the concatenation of spatial proximity, temporal proximity, word-overlap, and HoPS features. As explained in Section 5, we drop the tf-idf based feature owing to its poor performance. Instead, it serves as a useful baseline for comparison with word-overlap and HoPS based semantic features.

4.2 Learning Resources

What makes "understanding" the curriculum particularly challenging is the diversity of curriculum documents. The length of an LO text varies considerably, as does the scope of its underpinning concept. While a few LOs are independent, most are better understood in the context of their LRs, which elaborate on the dense information contained in the LO. In fact, Figure 3 depicts that most LOs are pithy, comprising fewer than 50 tokens (median token length = 6). In sharp contrast, LRs are lengthy documents with extensive detail (median token length = 578).

Whenever an LO is linked to an LR, we can obtain features from both of them. As mentioned in Section 3, LRs are well-curated slide decks. Inspired from Query Expansion (Vechtomova and Wang, 2006), we append the bold text from all slides of the linked LR to the LO text. Various semantic features, as detailed in Section 4.1.2, are then extracted from the expanded LO text and utilized in the proposed pipeline.

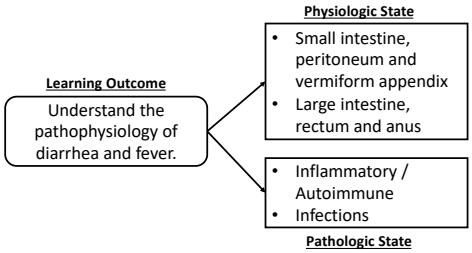

Figure 4: An LO annotated with relevant Physiologic and Pathologic states.

4.3 Expert Medical Codes (EMC)

To further enrich the feature space, we incorporate additional domain specific knowledge. For each LO, SMEs added a medical code indicating its 1) location or physiologic (Phys.) state, and/or 2) disease or pathologic (Patho.) state.

The physiologic code of an LO indicates the organ system it deals with, whereas the pathologic code specifies the nature of the disease or dysfunction covered by the LO. Experts hand-curated a total of 13 distinct physiologic states and 7 distinct pathologic states. Each LO could pertain to multiple physiologic and pathologic states, as depicted in Figure 4. Overlap between the EMCs of an LO pair is encoded as:

Categorical Similarity: Jaccard indices are obtained between 1) pathologic states and 2) physiologic states of the LOs in a pair, to ascertain overlap between their respective medical codes.

Semantic Similarity: While categorical similarity treats each physiologic and pathologic code as a distinct label, closer inspection reveals that there is inherent relatedness among the codes. For instance, the physiologic state "Pulmonary/Lung and Pleura" is closer in meaning to the state "Larynx, trachea, bronchi and alveoli", than it is to the state "Renal/Kidney". Thus, while comparing two LOs, we encode the word overlap of their respective medical codes as detailed in Eq. 1.

The final representation for the EMCs of an LO-pair is the concatenation of categorical similarity and semantic similarity features of the codes.

4.4 Cascaded Classification

A crucial aspect of this dataset is its extreme class imbalance. As Tables 2 and 3 demonstrate, LO pairs with `None` relationship vastly outnumber `Strong` or `Weak` pairs. This is to be expected, since the medical curriculum is extensive, and a

Confusion Matrix		Annotator 1			
		Strong	Weak	None	Total
Annotator 2	Strong	54	24	1	79
	Weak	38	64	25	127
	None	1	22	323	346
	Total	93	110	349	552
Macro-Average F1= 69.9, Accuracy= 79.9					

Table 2: Inter-Annotator agreement between two experts on the test set (note the substantial disagreement in annotating LO pairs as Strong and Weak).

Dataset	Strong	Weak	None	Total
Train	235 (14%)	344 (20%)	1,145 (66%)	1,724
Test	79 (14%)	127 (23%)	346 (63%)	552
Total	314 (14%)	471 (21%)	1,491 (65%)	2,276

Table 3: Train-Test splits. Note the class imbalance.

particular LO is likely to be related only to a small number of other LOs scattered in the curriculum.

While the priors of Strong and Weak classes are low, the risk in missclassifying them is high. Failure to identify a Strong LO pair is more detrimental than failure to identify a None pair. When we fail to recommend an LO strongly related to the one that a student is currently pursuing, it leads to a gap in their knowledge acquisition, whereas recommending an unrelated LO only leads to a degradation in user experience.

Additionally, we believe that the semantic gap between the three class labels is not identical. While it is relatively easier to distinguish None from Strong or Weak, the separation between Strong and Weak pairs is not as discernible. This is borne out further by the inter-annotator agreement in Table 2; for a large number of LO pairs, expert annotators disagree between Strong and Weak labels.

Aforementioned factors prompt us to split the 3-way classification task into two sequential binary classification tasks, as illustrated in Figure 2. The first classifier is trained on all input LO pairs, and classifies them as Related or Unrelated. In the next step, Related LO pairs are passed to the second classifier, which learns the degree of the relationship and further classifies them as Strong or Weak. LO pairs classified as Unrelated by the first classifier are directly labeled as None.

5 Experiments and Results

In all our experiments, we use NLTK for stopword removal and scikit-learn for the classifiers. We use $N = 20$ bins for HoPS features and set similarity threshold $\delta = 0.6$ for embedding-based features. We trained an SVM and Random Forest model for our task. Owing to space constraints and sub-par performance of the SVM, we report results for a Random Forest classifier with 100 estimators; all other parameters of the model are tuned using 5-fold cross validation on the training data. We use macro-F1 of the classifier on held out test data as our metric. Mean and standard deviations of macro-F1 are reported over 10 runs of the random forest. We use BioNLP(Chiu et al., 2016) word-embeddings.

For a subset of 552 LO-pairs, we obtain separate annotations from two SMEs. Inter-annotator performance (Table 2) on this held-out test set serves as a *skyline* for comparative evaluation. Owing to data-labeling constraints, only a subset of LOs could be linked to respective LRs by the SMEs. Similarly, tagging LOs with one of several possible physiologic/pathologic states entails significant cognitive engagement, and could be done only for a subset of LOs. For uniformity, we ensured that both subsets have a class label distribution identical to the total distribution in Table 3.

We perform three sets of experiments to 1) evaluate the effectiveness of the proposed approach, 2) evaluate the utility of LRs, and 3) evaluate utility of expert medical codes (EMC).

5.1 Evaluation

We compare five feature variants in an ablated study. Since the proposed approach stipulates curriculum and semantic features (CR+SM), we perform a comparison when individual curriculum (CR) or semantic features (SM) are used. To gauge the efficacy of tf-idf based features, experiments are performed using these features alone (TF), and along with curriculum features (CR+TF). For each feature variant, we contrast results obtained with a baseline 3-way monolithic classifier, and the proposed cascaded classifier. In the monolithic classifier, we ensure that the misclassification penalty for each class is inversely proportional to its frequency in the training data. This accounts for class imbalance, and ensures fair comparison against the cascaded classifier. Results of experiments are reported in Table 4. Our the key observations are :

Exact vs Embedding-based Features: Tf-idf features (TF) perform exact token matching which gets derailed whenever similar concepts are addressed differently (such as *myocardial* and *cardiac*). Instead, embedding-based features (SM) are more adept at capturing semantic relatedness as by construction, context vectors for related con-

Features	Classifiers	
	Baseline	Cascaded
CR	57.6±2.5	58.8±2.9
TF	43.1±0.9	49.8±1.2
CR+TF	53.2±1.3	55.9±2.0
SM	58.4±2.0	59.9±1.8
CR+SM	**63.6±1.1**	**66.1±2.3**
Inter-annotator agreement: 69.9		

Table 4: Macro-F1 (mean±std) values on the test set for two classifier variants and different features.

Features	Classifiers	
	Baseline	Cascaded
CR+SM without LR	63.3±1.5	65.9±1.3
CR+SM with LR	65.1±2.2	**67.2±1.7**
Inter-annotator agreement: 70.0		

Table 5: Macro-F1 (mean±std) on LR-linked test set.

cepts are closely located in the embedding space. Similarly, CR+SM outperforms CR+TF.

Importance of Feature Concatenation: A combination of both curriculum and semantic features (CR+SM) significantly outperforms their individual performance. Answering **RQ1**, we conclude that curriculum and semantics encode distinct aspects of an LO-pair's relatedness, and our system improves when information encoded in each feature class is jointly represented.

Effectiveness of Cascaded Classifier: For all feature combinations, the cascaded classifier outperforms the monolithic baseline. This supports our hypothesis that the decision boundary between `Related (Strong + Weak)` and `Unrelated` pairs is more discernible than the decision boundary between `Strong` and `Weak`.

For the rest of our experiments, we utilize CR+SM features with a cascaded classifier, since this combination yields best results, and approaches near human performance (refer Table 4). The proposed pipeline can now be used to establish LO relationships on the whole curriculum. This effectively circumvents the scale problem that manual annotation of all LO-pairs ($\sim$1 million) in the curriculum entails, while maintaining the accuracy of an expert.

5.2 Utility of Learning Resources

As reported in Table 5, it is clear that using LR text along with LO text improves LO-relationship extraction. This satisfactorily answers the question raised in **RQ2**. The dearth of adequate information and context in a concise LO poses a challenge for data-driven methods to ascertain semantic relatedness. LRs help plug this gap since they are more detailed and help expand the scope of both of our algorithms.

Features	Classifiers	
	Baseline	Cascaded
CR	65.8±2.3	68.1±2.1
SM	64.8±3.2	68.3±2.9
CR+SM	**70.5±2.4**	**72.6±2.0**
CR+SM+EMC	69.9±2.5	72.9± 2.3

Table 6: Macro-F1 (mean±std) values on ten random splits comparing the baseline, and inclusion of EMCs.

5.3 Utility of EMCs

Using features extracted from EMCs (detailed in Section 4.3), we compare the following combinations : curriculum (CR), semantic (SM), proposed concatenation of both (CR+SM), and subsequent concatenation with features from EMCs (CR+SM+EMC). Table 6 reports comparative results over 10 random 75-25% train-test splits.

We note that contrary to expectation, inclusion of EMC features (CR+SM+EMC) does *not* improve over CR+SM. We hypothesize that this may be because the classifier trained over CR+SM features learns an intermediate representation that correlates closely with the patho and physio states, thus their explicit inclusion provides no additional information to the classifier. While we may not know precisely what form the internal representation takes, it is interesting to note that our handcrafted features (CR+SM) and cascaded classifier design are both powerful enough to uncover underlying patterns of similarity between LOs. To answer **RQ3**, our approach does exploit latent patterns in the data.

6 Case Map Generation

The LO relationship extraction system can be applied to uncover LOs relevant to a virtual patient case (thus addressing **RQ4**). A virtual patient case describes a real-life scenario where a patient presents at the clinic with certain symptoms, and is administered specific tests. The medical student is expected to assume the role of a health-care professional and develop clinical skills such as making diagnoses and therapeutic decisions.

Figure 5 depicts part of a clinical case that has been annotated by SMEs. Crucial aspects of the case are highlighted as clinical factors, which may be symptoms (fever, hypotension, etc.) as well as diagnostic and screening tests. Each clinical factor is further linked to few pertinent *anchor* LOs. Successfully addressing a virtual patient case involves understanding these LOs, which may be

Clinical Case

28 year old male resident in Singapore presents with 6 day history of sweats, hot and cold spells, lethargy, headache, eye pain, epigastric pain and pain everywhere, persistent daily vomiting (4x/day), tiredness affecting work and dormitory life. No travel abroad in last 2 years. No contact with animals/ persons ill. He does not smoke and does not take alcohol. He does not think there are any illnesses in the family.

Blood tests showed NS1 dengue screen positive, Hb17.1, WCC 3.3, Platelets 20, Na 130, K 4.4, Urea 6.3, Cr 67, CXR clear. Examination showed vitals T38, HR120, BP 97/80, respiratory rate 20/min, oxygen saturations 98% on air. Malaria films were negative. Blood cultures were negative.

Patient deteriorated with nausea, hypotension, respiratory distress; and was transferred to ICU for supportive care including intubation and inotropic support. Bedside OGD showed 5 antral ulcers.

Clinical Factors

- Fever
- Abdominal Pain / GI Bleed
- Nausea / Vomiting
- Hypotension
- Diagnostic Tests and Screens

Figure 5: Annotated clinical case.

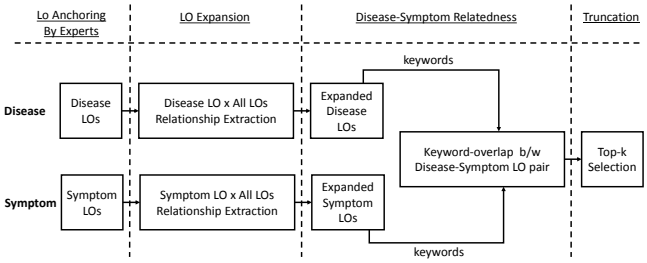

Figure 6: An approach to generate case map based on proposed LO-relationship extraction.

drawn from different years and disjoint sections of the curriculum. We aim to construct a *case map* that provides learners with a comprehensive view of the clinical case, in terms of its constituent LOs and their relationships. The map is envisioned as a graph, where nodes represent LOs, and edges establish relationships between them.

In attempting this, we encounter two primary challenges: 1) While SMEs can annotate a few anchor LOs, it is not feasible to manually enumerate all LOs related to the virtual patient case. This calls for an accurate LO-relationship extraction system that does not rely on expert intervention. 2) We must guarantee that these LOs are related within the context of the case. Since LOs by themselves do not provide enough textual content, we must look to LRs to ascertain whether LOs proposed by the system are appropriate in the context of the case at hand.

Given a disease, its symptoms, and diagnostic tests, we assume the availability of anchor LOs pertaining to each of them, and propose an approach outlined in Figure 6.

LO Expansion: The LO-relationship extraction system sequentially pairs an anchor LO with every LO in the curriculum, and classifies the relationship between them. For our purpose, we retain LOs that are labeled `Strong`, and disregard the rest. Thus, starting with a small set of anchor LOs, we obtain an expanded set of LOs that is strongly related to them.

Disease-Symptom Relatedness: We pair a disease-specific LO with a symptom (or diagnosis) specific LO, and measure the semantic overlap between their linked LRs. Higher the overlap, more relevant is the symptom (or diagnosis) LO to the disease LO. Thus, for each symptom (or diagnosis), their LOs are ranked by relevance to the disease LOs.

Truncation: The ranked list can be pruned to select the topmost k symptom (or diagnosis) LOs. Truncation ensures that for each symptom (or diagnosis), we select high-precision relationships with the disease specific LOs (characterized by overlap between their LRs). In the case map, this translates to at most k edges between a disease and each of its symptoms.

Figure 7 depicts the constructed case map for

LO ID	LO Text
LO_6704	Recall the clinical presentation and management of Dengue fever and Chikungunya infection
LO_4880	Describe how the presence of a viral infection may trigger off production of endogenous pyrogens leading to development of fever
LO_4881	Describe how bacterial infections may produce exogenous pyrogens resulting in the development of fever
LO_4882	Briefly describe how fever complements the immune response in infection
LO_6170	List the other abdominal organs that maybe responsible for abdominal pain
LO_6174	Explain the pathways controlling vomiting and nausea
LO_6175	Recall the use of vomiting patterns in differential diagnosis
LO_7793	Describe the role of relevant investigations for fever, including: Blood tests - hematology, chemistries, serology; Clinical samples - blood, respiratory, stool, urine, body fluids; Microbiology tests - cultures, PCR, serology; Imaging - Xrays, CT, MRI, ultrasound
LO_7794	Describe the role of relevant investigations for common infections, specific pathogens including dengue, malaria, typhoid, HIV, TB, MRSA
LO_7834	Define (and perform if relevant) appropriate resuscitation , immediate life support and acute management of: septic shock, neutropenic sepsis, dengue shock syndrome, severe malaria, acute bacterial meningitis

Table 7: Identifiers and text of the LOs that are part of the generated case graph in Figure 7 .

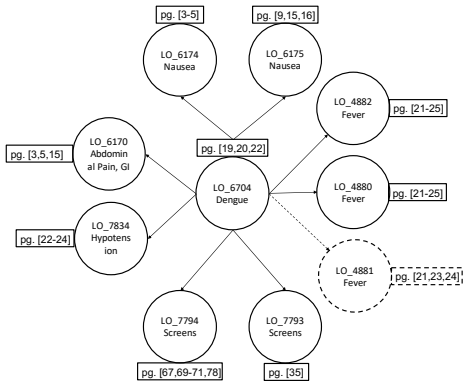

Figure 7: Case map extracted using proposed approach (see Table 7 for LO text corresponding to the LO IDs).

the dengue clinical case presented in Figure 5. Of the five clinical factors, four correspond to symptoms, namely : 1) Fever, 2) Abdominal Pain and GI bleed, 3) Nausea and Vomiting, and 4) Hypotension. The last factor corresponds to diagnostic tests and screens for dengue. We set $k = 3$, permitting at most 3 edges between dengue and each clinical factor. The generated case map was evaluated by an SME; one LO (LO_4881) is found to be spuriously a part of the map, whereas rest of the connections are deemed valid, thus establishing the efficacy of our approach.

7 Conclusion and Future Work

This work summarizes our effort to extract LO relationships using both semantic and curriculum cues. Owing to its human-level performance, our system serves as a reliable building block in constructing a case map from a virtual patient case.

Going forward, we would like to generate a concept map for all five years of the curriculum. We could then employ network analysis tools to uncover central LOs that drive most of the linkages. Secondly, we would like to *characterize* relationships between edges. Given a pair of related LOs, a simple characterization would be to assert if one of them is a pre-requisite to the other. Besides, we have observed that the classifier learns an intermediate representation that corresponds closely to EMCs. We could investigate if this can be harnessed to *predict* the states, thereby enriching curriculum metadata.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 252–263.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511.

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393. Association for Computational Linguistics.

Vincent Aleven, Amy Ogan, Octav Popescu, Cristen Torrey, and Kenneth Koedinger. 2004. Evaluating the effectiveness of a tutorial dialogue system for self-explanation. In *International Conference on Intelligent Tutoring Systems*, pages 443–454. Springer.

Yang Chen, Pierre-Henr Wuillemin, and Jean-Marc Labat. 2015. Discovering prerequisite structure of skills through probabilistic association rules mining. *International Educational Data Mining Society*.

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical NLP. In *Proceedings of the 15th workshop on biomedical natural language processing*, pages 166–174.

Julio Guerra, Yun Huang, Roya Hosseini, and Peter Brusilovsky. 2015. Graph analysis of student model networks. In *CEUR Workshop Proceedings*, volume 1446. University of Pittsburgh.

R.M. Harden. 1999. What is a spiral curriculum? *Medical Teacher*, 21(2):141–143. PMID: 21275727.

R.M. Harden. 2002. Learning outcomes and instructional objectives: is there a difference? *Medical Teacher*, 24(2):151–155. PMID: 12098434.

Rogers Jeffrey Leo John, Thomas S McTavish, and Rebecca J Passonneau. 2015. Semantic graphs for mathematics word problems based on mathematics terminology. In *EDM (Workshops)*.

Tanja Käser, Severin Klingler, Alexander Gerhard Schwing, and Markus Gross. 2014. Beyond knowledge tracing: Modeling skill topologies with bayesian networks. In *International Conference on Intelligent Tutoring Systems*, pages 188–198. Springer.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 133–142.

Jirí Rihák and Radek Pelánek. 2017. Measuring similarity of educational items using data on learners' performance. In *Proceedings of the 10th International Conference on Educational Data Mining (EDM)*, pages 16–23.

Swarnadeep Saha, Tejas I Dhamecha, Smit Marvaniya, Renuka Sindhgatta, and Bikram Sengupta. 2018. Sentence level or token level features for automatic short answer grading?: Use both. In *International Conference on Artificial Intelligence in Education*, pages 503–517. Springer.

Slavomir Stankov, Marko Rosić, Branko Žitko, and Ani Grubišić. 2008. Tex-sys model for building intelligent tutoring systems. *Computers & Education*, 51(3):1017–1036.

Kurt VanLehn. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221.

Olga Vechtomova and Ying Wang. 2006. A study of the effect of term proximity on query expansion. *Journal of Information Science*, 32(4):324–333.

Amal Zouaq and Roger Nkambou. 2008. Building domain ontologies from text for educational purposes. *IEEE Transactions on learning technologies*, 1(1):49–62.