Remediating textbook deficiencies by leveraging community question answers

Krishnendu Ghosh

Received: 2 November 2021 / Accepted: 2 February 2022 / Published online: 11 April 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract
The paper presents a method for recommending augmentations against conceptual gaps in textbooks. Question Answer (QA) pairs from community question-answering (cQA) forums are noted to offer precise and comprehensive illustrations of concepts. Our proposed method retrieves QA pairs for a target concept to suggest two types of augmentations: basic and supplementary. Basic augmentations are suggested for the concepts on which a textbook lacks fundamental references. We identified such deficiencies by employing a supervised machine learning-based approach trained on 12 features concerning the textbook’s discourse. Supplementary augmentations aiming for additional references are suggested for all the concepts. Retrieved QA pairs were filtered to ensure their comprehensiveness for the target students. The proposed augmentation system was deployed using a web-based interface. We collected 28 Indian textbooks and manually curated them to create gold standards for assessing our proposed system. Analyzing expert opinions and adopting an equivalent pretest-posttest setup for the students, the quality of these augmentations was quantified. We evaluated the usability of the interface from students’ responses. Both system and human-based evaluations indicated that the suggested augmentations addressed the concept-specific deficiency and provided additional materials to stimulate learning interest. The learning interface was easy-to-use and showcased these augmentations effectively.

Keywords Concept extraction · Deficiency diagnosis · Query generation · Question retrieval · Textbook augmentation

Krishnendu Ghosh
krishnendughosh@soa.ac.in

1 Department of Computer Science and Engineering, Institute of Technical Education and Research, Siksha ‘O’ Anusandhan (SOA) University, Bhubaneswar, India
1 Introduction

The institutional education model, as we know, has been carried out using a classroom-based environment for centuries. However, the last decade discerned a massive paradigm shift with the introduction of “e-learning,” offering quality education for everyone anywhere and anytime (Al-Qatawneh et al. 2019; Gelderblom et al. 2019). Although the e-learning systems kept changing their content structures (from traditional to multi-modal and customized) with time (Kempe and Grönlund 2019), they have always been favored due to their low-cost educational resources (Kopciewicz and Bougsiaa 2021; Nipa and Kermanshachi 2020). Even more, there has been no significant difference between the conventional classroom-based environment and self-study sessions supported by the e-learning paradigm as more and more users are getting along with this change (Leonard et al. 2021; Schuyten et al. 1999). During the recent Covid-19 pandemic, the schools remain closed across the globe (Kingsbury 2021). Consequently, the e-learning platforms continue enhancing the way they have been presenting content knowledge through various modalities (Ghosh et al. 2021; Mershad et al. 2019; Mitra and Gupta 2020; Singh et al. 2021). However, textbooks are still considered one of the primary conduits for catering basic knowledge forms to the students. The quality of these textbooks remains significant for catering primary knowledge forms to the students. The quality of these textbooks often suffers due to (i) the lack of detailed illustration of the concepts (Mohammad and Kumari 2007), (ii) improper sequencing of the concepts (Tyson-Bernstein 1988), and (iii) absence of adequate references (Mohammad and Kumari 2007). A substantial amount of studies to augment the textbooks have been undertaken to overcome these issues. Augmentations from these studies can be categorized into three classes: (i) bibliographic, (ii) component-based, and (iii) cQA-based. Bibliographic augmentations (e.g., Wikipedia articles or search-snippets) are suggested to offer additional references (Garcia-Gonzalez et al. 2016; Augustijn et al. 2018). Being elaborate in nature, these text-based augmentations address several related concepts instead of focusing on a particular one. Component-based augmentations are recommended in the form of a relevant image or video that illustrates a specific topical discourse (Agrawal et al. 2012b, 2014). However, such augmentations do not furnish any additional references for motivated students. cQA-based augmentations are provided in the form of QA pairs combining texts and relevant components (e.g., solved problems, figures, or real-life examples). They illustrate the target concepts using precise descriptions. Even so, they are reared with necessary components to understand them better and additional references that stimulate the students’ interests. QA pairs are also considered more comprehensible to the students than a formal textbook addressing their conceptual gaps (Singh et al. 2015). With this motivation, recent studies started augmenting textbooks using the cQA-based augmentations (Kumar and Chauhan 2019; Srba et al. 2019).
cQA platforms (like Yahoo! Answers\(^1\) or WikiAnswers\(^2\)) allow users from different backgrounds to satisfy their information need by browsing the historical archive of question-answer (QA) pairs and posing/answering direct questions. As the cQA forums gained popularity, their respective archives stacked a huge amount of question-answer pairs. These archives are now used to recommend relevant questions and, consequently, avoid the time-lag between asking a question and obtaining a personal response from the experts (Zhou et al. 2013). The popularity of the cQA platforms gradually led to their use in the education domains. There have been numerous efforts for developing educational cQAs in different scales (large-scale cQAs like Stack Exchange\(^3\), MedHelp\(^4\), Chegg\(^5\), Piazza\(^6\), or Brainly\(^7\) and small-scale cQAs (Aritajati and Narayanan 2013; Srba and Bielikova 2015)). Srba et al. (2019) analyzed the significance of such cQAs in the educational environments. These cQAs were developed as intervention-based forums offering little scope for the inactive students who hate participating. Consequently, there have been studies to improve the student experience and augment textbooks using the rich educational content offered by these cQAs (Uluyol and Agca 2012). Combining two related subjects (e.g., Geometry and Visual Arts) over similar cQA items, Schoevers et al. (2019) enriched the overall learning process. Kumar et al. (2019) proposed an approach for augmenting each sentence from the textbooks with cQA contents. However, these augmentations were recommended without identifying the conceptual deficiencies. Such augmentations often target an adequately addressed concept in the textbook but seldom enrich the deficient one. In the present work, we have proposed a system to suggest textbook augmentations in the form of question-answer (QA) pairs after examining the conceptual deficiencies.

The challenges in developing the proposed system have been:

1. **Extracting concepts:** With the availability of various annotation services, one could easily extract concepts from a textbook. However, the difficulty lay in extracting the key phrases which render domain-specific significance.

2. **Identifying deficiencies:** Concept-specific deficiencies had not been studied for Indian school-level textbooks. One needed to codify the possible deficiencies and model them over the textbook quality factors (Agrawal et al. 2012a).

3. **Retrieving augmentations:** Existing studies performed textbook augmentations without investigating concept deficiencies (Singh et al. 2015). QA pairs addressing the concept deficiencies could be retrieved by (i) creating useful queries, (ii) developing a potent question retrieval model, and (iii) filtering the augmentations to align them for the target students.

---

\(^1\) https://in.answers.yahoo.com  
\(^2\) https://wiki-answers.com  
\(^3\) https://stackexchange.com  
\(^4\) https://www.medhelp.org  
\(^5\) https://www.chegg.com/  
\(^6\) https://piazza.com  
\(^7\) https://brainly.in
The current work has proposed a textbook augmentation system that addresses concept-specific deficiencies present in the textbooks and furnishes additional references to stimulate learning interests. To diagnose the conceptual deficiencies, the current work investigated the factors which led to the deficiencies and modeled the diagnosis of deficiency as a supervised classification problem. Once the concepts had been diagnosed, queries were generated, and using a three-stage question retrieval approach; relevant QA pairs were retrieved from the popular community question-answering forum Stack Exchange. Stack Exchange is a chain of sites covering all possible educational fields. So, it is possible to locate relevant augmentation items from a single archive for different subjects and domains. The QA pairs from the Stack Exchange sites also come with detailed and organized metadata information. Such knowledge can be helpful to identify the validity of questions, acceptability of the answers, or expertise of the users and, in turn, retrieve quality augmentations items. Due to its archive size and organized metadata, we have considered Stack Exchange over any other cQAs. Finally, we deployed the augmentation system through a web-based user interface to present the augmented textbook contents to the students.

Considering the challenges mentioned earlier, the main contributions of this work have been as follows:

1. The current work recommended two types of augmentations: (i) basic and (ii) supplementary. Basic augmentations address the concepts that have been deficient in the textbook. On the other hand, supplementary augmentations offer additional references to stimulate learning interests. All the augmentations have been filtered to ensure that they are comprehensible to the target students.
2. Possible deficiencies in the school-level textbooks were surveyed, and the associated factors have been spotted. The proposed deficiency diagnosis module, modeled over these factors, automatically identifies these deficiencies.
3. The augmentation retrieval module has been designed by combining question retrieval models with suitable reranking and filtering techniques. While retrieval and reranking techniques ensured relevance of the augmentations, the filtering process guaranteed that they were understandable to the target students.
4. An easy-to-use web-based platform has been developed where students can easily access the augmentations while reading a learning material.
5. The proposed augmentation system has been assessed using a thorough evaluation plan. System-based evaluation tasks were performed for major modules of the proposed augmentation system. Quality of the augmentations and usability of the interface were assessed using human-based evaluation plans.

The paper has been organized as follows. Section 2 discussed the existent works related to the modules of the proposed system and has summarized the research gaps. Section 3 illustrated the architecture of our augmentation system while its
constituent modules have been elaborated in subsequent Sections 4-7. Implementation details of the interface were presented in Section 8. Section 9 presented the evaluation strategies and the system performance. Section 10 has concluded the present work with the directions for future work.

2 Related works

2.1 Related works on concepts extraction

Apart from the rule-based domain-dependent approaches (Wang et al. 2021), existing studies for extracting the set of phrases indicative of topical discourse of the study-materials are classified as supervised and unsupervised approaches.

2.1.1 Supervised concept extraction methods

One of the major supervised techniques is implemented as project KEA that employed Naive Bayes classifier to group the concepts based on their TF-IDF scores, and first occurrence (Witten et al. 2005). Nguyen improved the extraction veracity using positional information and a set of morphological features (Nguyen and Kan 2007). Recent approaches explored relevant features to model the concepts and their keyphraseness over the training data to extract the concepts. However, supervised techniques have never been a quality option for developing domain-independent systems.

2.1.2 Unsupervised concept extraction methods

There have been several techniques for extracting concepts in an unsupervised approach. One approach was based on different statistics-based measures, e.g., TF-IDF scores, context, and span of the candidate concepts (Campos et al. 2018; El-Beltagy and Rafea 2009). With the introduction of Wikipedia, several Wikipedia-based properties (e.g., Wikipedia-based keyphraseness (Medelyan et al. 2009), Wikipedia links, Wikipedia-based IDF, phrase embedding, keyphrase- ness, and title overlap (Papagiannopoulou and Tsoumakas 2020)) were employed to improve these systems. The second approach was graph-clustering methods, where a graph is generated using the candidate concepts as nodes and their relations as edges. This graph is clustered, and the representative concepts from each cluster are drawn out to forge the set of key concepts. One of the earlier graph-clustering approaches named TextRank modeled the relations between the concepts based on their binary co-occurrence measure in a $M$-word frame (Mihalcea and Tarau 2004). This method has been boosted where the edges were weighted based on the amount of co-occurrence in SingleRank (Wan and Xiao 2008), and RAKE (Rose et al. 2010). The third approach employed embedding-based techniques for extracting concepts. Bennani-Smires et al. (2018) designed EmbedRank using two different embeddings (Doc2Vec and Sent2Vec) for representing a document and the underlying candidate concepts. The candidature of a concept
was finalized based on the distance between these embeddings. While the first two approaches suffer due to their expensive training data collection and feature extraction steps, embedding-based methods are considered easier to implement. However, being one of the fundamental and well-studied natural language processing tasks, there lies little scope for improving the performance of a concept extraction task. Such tasks are often outsourced to several API-based services (e.g., DBpedia or Tagme) that combine the available concept extraction features. The linked Wikipedia articles also offer supplementary details of the concepts. Consequently, recent works started banking upon DBpedia, Tagme, or similar tools for concept extraction.

2.2 Related works on diagnosing deficiencies

Substantial amount of studies have been undertaken for (i) determining the factors affecting the overall quality of the textbooks (Pawlowski et al. 2007; Woodward et al. 2013), (ii) proper sequencing of the curriculum (Chambliss and Calfee 1998), and (iii) creating subject-wise guidelines for textbook creation (Crossley and Murby 1994). Summarizing these studies, textbook deficiencies can be categorized as style-centric and concept-specific.

Style-centric deficiencies depend on two aspects. The first aspect concerns the individual contrasts (e.g., intellectual capacity, reading competence, prior experiences, and personal intrigue). The second set relates to the readability of the materials (e.g., format, organization, or writing style) (Kieras and Dechert 1985). As the first category does not directly associate with textbooks, we analyze the latter aspect using four linguistic variables, namely, vocabulary load (Lockheed and Hanushek 1988), sentence structure, idea density (Asheim 1958), and human interest (Agrawal et al. 2011a). Such factors have been integrated into several formulas (Flesch 1948; Dale and Chall 1948; Mc Laughlin 1969; Coleman and Liau 1975). The current work focuses on identifying the concept-specific deficiencies and, therefore, employs simple linguistic features (average sentence length, average word length, and average word familiarity).

Good-quality textbooks are typically organized into chapters and subdivided into sections. They should be designed such that each section (i) precisely illustrates a few related concepts, (ii) explains the topics using necessary figures, examples, or solved problems (Chambliss and Calfee 1998), (iii) discusses adequate and relevant references, (iv) follows a sequence so that, a new concept is not encountered before explaining all of its prerequisite concepts (Anderson et al. 1984), and (v) is arranged in a way where related concepts should be discussed in consecutive sections without creating memory-burden in students (Agrawal et al. 2012a). Investigating the issues above, Agrawal et al. (2011a; b; 2012b) recommended enrichments for textbook sections if found deficient, employing linguistic factors like readability, different forms of syntactic complexity, and dispersion between concepts. However, proceeding from the existing literature, concept-specific features need to be determined to diagnose concept-specific deficiencies.
2.3 Related works on suggesting augmentations

The task of suggesting augmentations comprises three different alleys. The first approach surveyed how different factors are influenced using different learning resources in a learning process. The second one retrieves the augmentations from the freely available Web resources and offers them to the students without any intervention. The second approach designs its own educational cQA, where the students require to participate in active knowledge-sharing tasks.

Adesope et al. (2017) compared concept maps, refutations, and expository text in diminishing concept-gaps among the students. Purgina et al. (2020) established the effectiveness of gamification by developing a game “WordBricks” for second-language acquisition. Klippel et al. (2019) showed how a virtual field trip would cut costs yet could achieve a better learning experience and actual lab scores. Wang and Xing (2019) designed a structural model to identify how elementary students’ learning gains are affected based on external factors like self-efficacy, learning motivation, learning strategies, and parent education level.

Research works for textbook augmentation have been initiated with the introduction of Web resources and cQAs. Such augmentations are primarily composed of texts, relevant components (such as images, videos, solved problems), or QA pairs. Koć-Januchta et al. (2020) introduced questions relevant to the topics discussed in an AI-enriched e-book to help in retaining and analyzing the concerned topics. In a similar study, Alsalem (2018) demonstrated how meta-cognitive strategies could be used in an e-book to develop better reading comprehension skills among the DHH (deaf and hard of hearing) students. A few studies also incorporated AR (Augmented Reality) techniques for specific learning tasks - learning atoms’ and molecules’ reactions Ewais and Troyer (2019) and learning biology concepts (Weng et al. 2020). Most of these works suggested augmentations without investigating the deficiencies of the textbooks. Such augmentations often enrich the concepts even though they are precisely explained, but seldom address the deficient concepts (Singh et al. 2015). Although they claim to provide better learning, it is prudent not to provide definitional augmentations for the concepts which have been discussed sufficiently in the textbook.

Growing popularity of cQAs motivated the recent works on designing and developing educational cQAs. Apart from offering a platform for knowledge-sharing among the students, such cQAs work towards routing questions to the best experts ([Macina et al. 2017]), balancing load between the experts (Babinec and Srba 2017) and refining the answers based on content-quality (Choi et al. 2015; Le et al. 2016). Social networking sites triggered Brainly9 to incorporate real-time chats and discussion forums. In subsequent systems like Piazza10 or Askalot11, teachers also take part in answering queries and examining contents. While such systems are beneficial

---

9 https://brainly.in
10 https://piazza.com
11 https://askalot.fiit.stuba.sk/demo
for active students, passive students remain deprived of the automatically recommended augmentations (Srba et al. 2019).

Limitations about the tasks of diagnosing deficiencies and recommending augmentations can be enumerated from the surveyed literature as follows.

- Concept-specific deficiencies have not been surveyed for Indian school-level textbooks. Accordingly, the factors responsible for such deficiencies have also not been investigated systematically.
- None of the existing textbook augmentation systems have diagnosed concept-specific deficiencies. Subsequently, the recommended augmentations have failed to address the deficient concepts precisely but have often enriched the adequately illustrated concepts.

3 Proposed Architecture

The proposed augmentation system has been realized using four major modules: (i) concept extraction, (ii) deficiency diagnosis, (iii) query generation, and (iv) textbook augmentation, as outlined in Fig. 1.
Let us consider, a textbook \( T \) for grade level \( GL \) is organized over a set of sections \( S = \{ s_1, s_2, \ldots, s_m \} \). Each of these sections \( s_i \) is illustrated using a set of concepts \( C(s_i) = \{ c_1, c_2, \ldots, c_n \} \). From each section \( s_i \), first, the concepts \( C(s_i) \) were extracted. Each concept \( c_j \in C(s_i) \) from section \( s_i \) were diagnosed to classify its deficiency \( \text{Def}(c_j) : C \rightarrow D \), where \( D \) is the set of deficiency types or classes. For concept \( c_j \), the subsequent module generated basic query \( BQ(c_j) \), addressing deficiency \( \text{Def}(c_j) \) and supplementary query \( SQ(c_j) \), that provides additional references on \( c_j \). These queries were formed using concept \( c_j \), its deficiency type \( \text{Def}(c_j) \) and neighboring context \( \text{Con}(c_j) \). Querying the input question archive \( QA \) with query \( BQ(c_j) \) (or \( SQ(c_j) \)), a set of final set augmentations \( BA(c_j) \) (or \( SA(c_j) \) ) was retrieved. In the subsequent sections, the input-output specifications and the implementation details for each of the constituent modules are presented.

4 Concept Extraction

Glossary, provided in a textbook, is the most reliable resource to extract the domain-specific concept phrases (Mulvany 2009). In the absence of these expert-administered indexes for most of the textbooks, the present work proposed an alternate approach to extract the set of key concepts present \( C(s_i) = \{ c_1, c_2, \ldots, c_n \} \) for each of the sections \( s_i \in S \) of a grade level \( GL \) textbook \( T \) organized over a set of sections \( S = \{ s_1, s_2, \ldots, s_m \} \).

To extract meaningful relevant concepts (or, entities) the current work extracted concepts using: (i) linguistic patterns (Agrawal et al. 2011a, b), (ii) DBpedia spotlight tool (Daiber et al. 2013) and (iii) glossary terms. Phrases with the linguistic pattern of \( A^*N^+ \) (where \( A \) represents an adjective and \( N \) a noun), commonly considered as concepts (Toutanova et al. 2003), was first extracted from the textbook sections. The list of concepts was further populated with the concepts that are annotated by the DBpedia tool and mentioned in expert-labeled glossary terms. Concepts exactly matching the title of a Wikipedia article were only considered to generate the final list of concepts. The decision to use Wikipedia for validating the concepts was taken by considering its coverage over the topics in basic subjects. However, that did not rule out the possibility of discarding a valid concept not present in Wikipedia.

5 Deficiency diagnosis

This module identifies the deficiency \( \text{Def}(c_j) : C \rightarrow D \) for a target concept \( c_j \in C(s_i) \) from textbook section \( s_i \). To introduce the deficiency types, let us first consider the properties of the quality textbooks.
5.1 Properties of Quality Textbooks

The properties of good quality textbooks can be listed as:

1. **Readability**: A good-quality textbook should avoid the use of long sentences, complex sentence structure, and clauses. With reduced syntactic and word-level complexity, readability gets improved (Gray and Leary 1935; DuBay 2004).

2. **Focus**: Each textbook section should explain a set of concepts that are related to each other (Clark et al. 2011).

3. **Unity**: Instead of being described across various sections, a concept should be discussed in a particular section (considered as the key section for the concept) (Chambliss and Calfee 1998).

4. **Sequentiality**: Concepts should be explained sequentially where (i) before introducing a new concept, all of its prerequisite concepts are explained, and (ii) all its related concepts are discussed in consecutive sections (Paas et al. 2003).
5. Relevant Components: Explaining concepts is easier when relevant figures, examples, or solved problems are present.

5.2 Deficiency classes

Based on the properties mentioned above, we associate a concept with one of the following deficiencies:

1. Focus-Deficiency (D1): Considering the property “Focus”, concepts of a textbook section should be related to each other. However, if a concept is sparsely related to the other concepts that, on the other hand, share substantial accordance, it is considered as focus-deficient. Relatedness between concepts is commonly presented using concept graphs where nodes are made of concepts and edges represent their
relatedness. Figure 2 offers a concept graph formed with concepts from a textbook section. In this section, the concept “momentum” is a D1-deficient concept as it shares relatively less connectivity with the other concepts in terms of relatedness. However, this relatedness can be increased in the presence of concepts “impulse,” “conservation of momentum,” “elastic collision,” or “inelastic collision.” Therefore, focus deficiency associated with concept “momentum” can be remediated by augmenting this section with mentioned missing links.

2. Sequence-Deficiency (D2): Introducing a concept without mentioning its prerequisite concepts may create unnecessary memory-burden among the students (Agrawal et al. 2012a). A concept discussed before or after the section where it should ideally be discussed is considered sequence-deficient. The sequence of the concepts from a textbook section is depicted in Fig. 3.

3. Component-Deficiency (D3): Complex concepts are meant to be discussed in detail with essential figures, examples, and relevant problems. In the absence of such components, students often find it difficult to assimilate the associated concepts. Visualizing the concepts “compression” and “rarefaction” is relatively easier when accompanied by an image, as shown in Fig. 4.

4. No deficiency (D4): A concept not having any of the deficiencies as mentioned earlier is termed as not deficient.

Deficient concepts can be identified by pointing out several aspects (e.g., related concepts, prerequisite concepts, dependent concepts, or relevant components). Manual annotation for these aspects being unfeasible; the current work relied on the conceptual structure of Wikipedia. Considering $A(c)$ as a Wikipedia article associated with concept $c$, we defined the relevant aspects as follows.

**Definition 1** Related concepts $R(c)$: The set of concepts related to a concept $c$ is represented as $R(c) = \{ c_k \}$ where the Wikipedia articles $A(c)$ and $A(c_k)$ are linked (by in-links or out-links).

**Definition 2** Prerequisite concepts $P(c)$: The set of concepts prerequisite to a concept $c$ is represented as $P(c) = \{ c_k \}$ if the article $A(c)$ has in-links from each Wikipedia article $A(c_k)$.

**Definition 3** Dependent concepts $D(c)$: The set of concepts dependent on a concept $c$ is represented as $D(c) = \{ c_k \}$ if the article $A(c)$ has out-links to each Wikipedia article $A(c_k)$.

**Definition 4** Relevant components $RC(c)$: A component is considered relevant for a concept $c$, if the concept or any of its related concepts are present in the title/caption of the component. Accordingly, the relevant components $RC(c)$ is defined as $\{ com \mid |R(c) \cap Title(com)| > 1 \}$ where $Title(com)$ is the set of concepts present in the title/caption for a component $com$.  

 Springer
Definition 5 Concept graph $G(C)$: A concept-graph ($G(C) = \langle N, E \rangle$) is a graph where $N$ and $E$ stand for the set of nodes and edges, respectively. $N$ models the set of concepts $C$. An edge $e_{ij} \in E$ between concepts $c_i$ and $c_j$ exists, if Wikipedia article $A(c_i)$ and $A(c_j)$ are connected (by in-links or out-links). Presence of single or multiple links between the articles is treated equally as the significance of the links cannot be determined without an external knowledge-base.

Definition 6 Section-based distance $dis(c_i, c_j)$: Let $s(i)$ and $s(j)$ are the key sections of the concepts $c_i$ and $c_j$ respectively. For a pair of concepts $c_i$ and $c_j$, the section-based distance $dis(c_i, c_j)$ is defined as the number of sections between $s(i)$ and $s(j)$ and is formulated as $dis(c_i, c_j) = 1 + |in(s(c_i)) - in(s(c_j))|$, where $in(s)$ is the index of the section $s$.

The heuristics-based annotation (based on definition 1-4) of different concepts calls for evaluating the quality of annotation. Consequently, the concepts that had been annotated using the proposed heuristics were manually inspected. The result of this manual inspection is reported using standard accuracy measure in Table 1. As the quality of these heuristics for predicting concept aspects was satisfactory, the conceptual structure of Wikipedia has been used to annotate the concepts with different aspects.

| Aspects           | Accuracy |
|-------------------|----------|
| Related concepts  | 0.77     |
| Prerequisite concepts | 0.79    |
| Dependent concepts | 0.71     |

5.3 Classification model and feature space

5.3.1 Classification model

Banking upon the performant Wikipedia-hierarchy and various aspects of concepts, we proposed a supervised classification-based approach for diagnosing deficiency. This approach learned a mapping function: $Def(c) : C \rightarrow \{D1, D2, D3, D4\}$ from a dataset consisting of ($\text{Concept, D1/D2/D3/D4}$) items collected from textbooks. The Support Vector Machine (SVM) with an RBF kernel, implemented using libSVM python package (Chang and Lin 2011) was used as a classifier. The kernel-type and the optimal parameters were determined using 10-fold cross-validation with validation and test set. The features representing different aspects of a concept/section and its deficiency are of two major classes: section-specific and concept-specific. Section-specific features, having the same values for all concepts from a section, are considered to meet better “Readability” and “Focus” across sections. Other properties of good-quality textbooks are reflected using
concept-specific features. For deficiency diagnosis, we employed four section-specific and eight concept-specific features, as mentioned in Fig. 5 and defined below.

5.3.2 Section-specific Features

Four section specific features considered in this work are:

1. **Average sentence length**: It is determined by averaging the number of words present in the sentences from the concerned textbook section.
2. **Average word length**: This is measured by averaging the number of letters present in each word from the concerned textbook section.
3. **Average word familiarity**: This measure is the normalized proportion of the words that are generally considered uncommon. This feature is often termed also as lexical diversity.

\[
WF(s) = \frac{|\{word_i | word_i \in DC\}|}{|Word(s)|} \tag{1}
\]
Word(s) is the set of words (allowing repetitions) in textbook section s and DC is the set of familiar words from Dale long list (Dale and Chall 1948).

4. **Dispersion between concepts:** According to property “Focus” Agrawal et al. (2012a), understanding a textbook section discussing related concepts is easier than one comprising many unrelated concepts. This premonition has been formulated as a measure of dispersion between the concepts of a textbook section following (Agrawal et al. 2011a). Dispersion measures the fraction of the unrelated concept pairs from the set of concepts C(s) present in textbook section s. Agrawal et al. (2011a) generated a concept-graph for the concepts C(s) with E edges and determined the fraction as follows:

$$CD(s) = 1 - \frac{|E|}{|C(s)|(|C(s)| - 1)}$$

Dispersion between concepts of a section was formally determined using Algorithm 1 following (Agrawal et al. 2011a).

**Algorithm 1 Determining conceptual dispersion CD(s) for section s.**

**Input:** Set of concepts C(s) for section s  
**Output:** Conceptual dispersion CD(s) of section s  
1. Let L denote the set of links (c_m, c_n) where Wikipedia articles A(c_m) and A(c_n) respectively associated to concepts c_m and c_n are linked.  
2. Generate a directed graph G = {C, E} where E is the set of edges defined as \{(c_m, c_n)|c_m, c_n \in C(s) \land (c_m, c_n) \in L\}.  
3. Compute conceptual dispersion CD(s) of section s using formula: $1 - \frac{|E|}{|C(s)|(|C(s)| - 1)}$

Dispersion value ranges between 0 and 1, where 0 relates to a section where each concept-pairs are related, and 1 corresponds to a section with mutually independent concepts.

5.3.3 **Concept-specific Features**

The concept specific features used in this work are:

1. **Related concepts absent count:** Number of related concepts for concept c absent in section s is determined as |R(c) − C(s)|.
2. **Prerequisite concepts absent count:** Number of prerequisite concepts for concept c absent in section s is determined as |P(c) − C(s)|.
3. **Dependent concepts absent count:** Number of dependent concepts for concept c absent in section s is represented similarly by |D(c) − C(s)|.
4. **Concept significance score:** The significance score measures the importance of a concept c in section s to perceive other concepts from the textbook. The significance score $sig(c, s) = freq(c, s) \ast |R_s(c)|$ where $R_s(c) = R(c) \cap C(s)$ Agrawal et al. (2012a).
5. **Key section count:** The number of key section for a concept c is defined as: $K_c = \{ s_k | s_k = \max_{s \in S} \{ sig(c, s) \} \}$
6. **Average section distance for related concepts**: Considering that the related concepts are to be discussed in the same or consecutive sections, section distance between two related concepts should be lower. Based on the premise, the average section distance \((ASD_R(c))\) for a concept \(c\) as:

\[
ASD_R(c) = \frac{\sum_{c_i \in R(c)} dis(c_i, c)}{|R(c)|}
\]  

(3)

Let us consider, a concept \(c\) is located in Section 1. Concepts, related to \(c\) are \(c_1\) and \(c_2\) from Section 2, and \(c_3\) from Section 4. So, \(R(c) = c_1, c_2, c_3\). Figure 6 illustrates how to calculate the average section distance between these related concepts. As \(c_1\) and \(c_2\) are located in the section next to the section where \(c\) is located, their distance is \(dis(c, c_1) = dis(c, c_2) = 1\). Similarly, \(dis(c, c_3) = 3\) and finally, \(ASD_R(c)\) becomes \(5/3\).

7. **Average section distance for prerequisites**: This feature estimates whether or not the prerequisites are discussed before a concept. For concept \(c\), average section distance between prerequisite concepts \((ASD_P(c))\) is defined as:

\[
ASD_P(c) = \frac{\sum_{c_i \in P(c)} dis(c_i, c)}{|P(c)|}
\]  

(4)

8. **Relevant component count**: Understanding a concept is relatively easier with a higher number of components illustrating the target concept. The number of components, relevant to concept \(c\) is defined as \(|RC(c)|\).
6 Query generation

To retrieve the basic and supplementary augmentations from cQA archive QA for a concept $c_j$, two keyword-based queries were generated. The basic query $BQ(c_j)$ was created by combining a concept $c_j$, their context $Con(c_j)$ and deficiency type $Def(c_j)$. Supplementary queries $SQ(c_j)$ were formed using the concept $c_j$ and its context $Con(c_j)$.

6.1 Basic Query Generation

The generation of the basic query was done by choosing context $Con(c_j)$ as follows:

- For Focus-Deficiency (D1), the context was formed using all the concepts related to the concerned concept.
- For Sequence-Deficiency (D2), the context was formed using the prerequisite and dependent concepts.
- For Component-Deficiency (D3), the context was created by using all the concepts that (i) have been related to the concerned concept but (ii) have no relevant components in the present textbook section.

Consequently, the basic query $BQ(c_j)$ for concept $c_j$ was formulated as $\langle c_j, Con(c_j) \rangle$ where $Con(c_j)$ is defined as:

$$Con(c_j) = \begin{cases} R(c_j) & \text{if } Def(c_j) = D1 \\ P(c_j) \cup D(c_j) & \text{if } Def(c_j) = D2 \\ \{c_k \in R(c_j) \mid |RC(c_k)| = 0\} & \text{if } Def(c_j) = D3 \end{cases}$$

(5)

6.2 Supplementary query generation

Considering the associations for a concept to be reflected from its related concepts, context $Con(c_j)$ for concept $c_j$ was formed using all the concepts present in $R(c_j)$.

$$SQ(c_j) = \langle c_j, R(c_j) \rangle \text{ if } Def(c_j) = D4$$

(6)

7 Textbook augmentation

This textbook augmentation module consists of three sub-modules: (i) Augmentation Item Retrieval, (ii) Structural Re-ranking, and (iii) Filtering Augmentations. In order to retrieve a ranked list of QA pairs from a QA archive, the proposed system employed a question retrieval model following (Ghosh et al. 2017) to determine the similarities between the ⟨query-QA⟩ pairs. Instead of overloading with numerous
augmentations, the current application aims to return highly relevant augmentations, easily comprehensible to the target students. Augmentation items retrieved initially were reranked based on their relevance with the query. The final list of augmentations was returned by filtering these reranked augmentations with the associated grade level.

### 7.1 Augmentation retrieval

Following (Ghosh et al. 2017), the current application modeled a similarity function between a query $q$ and QA pairs $Q_i \in QA$. This similarity function, devised over three major categories of features, returns a ranked list of augmentations $Q_{init}$. Each of these retrieved augmentations $Q_i \in Q_{init}$ obtains a score $IS_{Q_i}(q)$. The features used for learning the similarity function are categorized as follows:

#### 7.1.1 Lexical features

Features from this class compute the word level similarity between a query and QA pair. Following lexical features were used in the current work.

1. **Word n-gram overlap ($n = 1, 2$ and $3$):** It is the ratio of the number of overlapping n-grams between the $\langle$query-QA$\rangle$ pair and the number of n-grams present in QA.

2. **BM25 score:** This feature approximates similarity between a $\langle$query-QA$\rangle$ pair considering: term-frequency, inverse document frequency and length of the documents (Manning et al. 2010).

3. **Cosine similarity:** This feature is quantified as the cosine similarity to measure the lexical similarity between a query-QA pair based on word n-grams ($n = 1, 2$, and $3$) scores of words.

#### 7.1.2 Syntactic features

These features are computed from the syntactic structure of the sentences:

1. **Noun overlap:** It is the count of nouns present in both the query and candidate QA pair, normalized by the number of nouns present in the QA pair.

2. **Verb overlap:** It is computed by normalizing the overlap counts of verbs.

3. **Dependence-pair overlap:** Dependence-pairs present a set of grammatical dependency relationships between the words within a sentence (Chen and Manning 2014). For example, “Alexander” is considered as a passive nominal subject for the verb “impressed” in the sentence “Alexander was impressed by Porus.” Such a relationship is represented as a dependence-pair “nsubjpass(impressed, Alexander).” The count of common dependence pairs between a query-QA pair normalized by the number of dependence pairs present in the QA pair is proposed as a syntactic feature. The spaCy python package obtained these dependence pairs using Stanford parser (Chen et al. 2014).
4. **Named-entity overlap:** Named-entities refer to real-world objects such as persons, locations, organizations, or products. A normalized count of the common named entities between a query-QA pair is used as another feature. We used Stanford coreNLP python APIs to determine named entities.

### 7.1.3 Semantic features

Aiming to match the latent conceptual meaning, the following semantic features were used in the current application.

1. **Word alignments:** A query that is similar to a QA pair often tends to show significant word-level alignments. The similarity between \( (q; QA) \) pair can be quantified by aggregating alignment probabilities of constituent word/phrase pairs, one taken from \( q \) and the other from \( QA \). Using a parallel corpus of 100,000 similar QA pairs collected from Stack Exchange on a word-alignment model using Giza++\(^{12}\), phrase pair alignment probabilities had been determined. Finally, the similarities between the query and QA pairs were determined with statistical machine translation using IBM model-1.

2. **Common frames:** Frames present a schema to state particular relationships between different constituents in a sentence. For example, in the sentence “Alexander was impressed by Porus,” the relationship between “Alexander” and “Porus” can be affirmed using a frame “Impression.” While a sentence is represented differently, these different representations may often contain similar frames. The normalized count of the common frames between query and question has been considered as the feature “common frames”. These frames are obtained using Sling python API using FrameNet frames (Baker et al. 1998).

### 7.1.4 Question similarity model

The similarity function, parameterized over the features mentioned earlier, was modeled using a deep neural network (DNN), comprised of one input layer, two hidden layers, and an output layer following (Ghosh et al. 2017). The features were used in the input layer, and the output layer finally denoted the similarity. The weights were initialized in a pre-training step using greedy layer-wise Contrastive Divergence technique (Carreira-Perpinan and Hinton 2005) and were fine-tuned by standard backpropagation technique discriminatively.

### 7.2 Structural re-ranking

To promote the QA pairs from the set of initially retrieved augmentations \( Q_{\text{init}} \), a structural reranking technique was proposed in the present work based on the following assumption from Kurland and Lee (2005).

---

\(^{12}\) [http://www.statmt.org/moses/giza/GIZA++.html](http://www.statmt.org/moses/giza/GIZA++.html)
**Assumption 1** A QA pair \( (Q_i) \) should be ranked higher for a query \( (q) \) if it is similar to most of the other QA pairs retrieved for the same query.

To illustrate, consider a case where due to the absence of specific terms in \( q \), a QA pair \( Q_1 \) gets a lower similarity score where QA pairs \( Q_2 \) and \( Q_3 \) are scored higher. Considering the terms from the top-ranked QA pairs significant, \( Q_1 \)'s score can be modified based on the extent \( Q_1 \) matches with top-ranked QA pairs \( Q_2 \) and \( Q_3 \). Similarity with the other QA pairs, therefore, provides support to re-assess the rank for the associated QA pair using three steps: (i) Support Graph Generation (ii) Cumulative Support Determination, and (iii) Final Scoring.

### 7.2.1 Support graph generation

The support relationship between two QA pairs was determined based on their similarity value. Stating formally, a QA pair \( Q_i \) is assumed to receive support from another pair \( Q_j \in Q_{init} \) if \( Q_i \) is retrieved against a query \( Q_j \). The support relationships between all QA pairs from \( Q_{init} \) were encoded using a support graph.

**Definition 7** Support graph: A support graph is a directed and weighted graph \( G = (V, E) \). Each QA pair from \( Q_{init} \) is represented by a node \( v \in V \). A directed edge \( v[Q_i] \rightarrow v[Q_j] \) in \( E \) represents the support \( Q_j \) receives from \( Q_i \), quantified by a weight function \( wt(v[Q_i] \rightarrow v[Q_j]) \) defined in Definition 8.

**Definition 8** Edge weight function \( wt(\cdot) \): Consider that, \( Top_\alpha(Q_j) \) represents the set of top \( \alpha \) QA pairs \( \{Q_i \mid Q_i \in Q_{init} \setminus \{Q_j\} \} \) for query \( Q_j \). The support \( Q_j \) receives from \( Q_i \) is computed as follows:

\[
wt(v[Q_i] \rightarrow v[Q_j]) = \begin{cases} 
IS_{Q_i}(Q_j) & \text{if } Q_i \in Top_\alpha(Q_j), \\
0 & \text{otherwise}
\end{cases}
\]  

(7)

The construction of supports graph started by creating a set of nodes, which associated \( Q_{init} \). For each QA pair \( Q_i \), the construction strategy identified top \( \alpha \) similar questions in \( Q_{init} \setminus \{Q_j\} \) and established edges between each respective pair. This process was conducted iteratively for all the QA pairs in \( Q_{init} \). There might be cases where an insignificant amount of support has been propagated to a target pair from the others. Consequently, QA pairs that had achieved an average score in the first phase might suffer a hefty penalty. The current application proposed the Pagerank smoothing technique (Brin and Page 1998) to compensate for such discrepancy. Accordingly, a graph \( G \) was smoothed to \( G^{[\lambda]} \) with a smoothing parameter \( \lambda \in [0, 1] \). \( G^{[\lambda]} = (V, E^{[\lambda]}) \) is formed with smoothed edge weights (\( wt^{[\lambda]} \)) defined for every \( Q_i, Q_j \in Q_{init} \) below.

\[
wt^{[\lambda]}(v[Q_i] \rightarrow v[Q_j]) = (1 - \lambda) \frac{1}{|Q_{init}|} + \lambda \frac{wt(v[Q_i] \rightarrow v[Q_j])}{\sum_{Q_j' \in Q_{init}} wt(v[Q_i] \rightarrow v[Q_j'])}
\]  

(8)
7.2.2 Cumulative support determination

Amount of support QA pair $Q_i$ propagates to QA pair $Q_j$ depends on the support $Q_i$ acquires from others. Hence, cumulative support ($CSR$) is determined as:

$$CSR(Q_j, G^{[d]}) = \sum_{Q_i \in Q_{init}} w^{[d]}(v[Q_i] \rightarrow v[Q_j]).CSR(Q_i, G^{[d]})$$  \hspace{1cm} (9)

7.2.3 Final scoring

For a query-QA pair $(q, Q_j)$, the final score of $Q_j$ was modeled as a function of initial retrieval score $IS_{Q_j}(q)$ and cumulative support score $CSR$ where the associated parameters $\alpha$ and $\lambda$ had been learned to achieve better retrieval performance. Reranking the QA pairs $Q_j$ according to their score $SR_R(Q_j, q)$ against query $q$, a list of QA pairs $Q_{rerank}$ was returned.

$$SR_R(Q_j, q) = CSR(Q_j, G^{[d]}) \times IS_{Q_j}(q)$$  \hspace{1cm} (10)

7.3 Filtering augmentations

The set of reranked augmentation items $Q_{rerank}$ was filtered based on their suitability to the target population or population of a target grade level. The present work depended on the following assumptions to determine whether an augmentation relates to a target grade level $GL$ or not.

**Assumption 2** An augmentation is considered comprehensible to a student, if she can understand the concepts present in the augmentation.

**Assumption 3** A student from a particular grade level can understand a concept if the concerned concept is mentioned in a textbook prescribed for the associated grade level or lower.

Let $C_{GL}$ represents the set of concepts that appear in the index of the books for grade-level $GL$. The grade level for a concept is, therefore, determined as the grade level where it is first introduced:

$$Grade(c_j) = \arg \min_{GL} [c_j \in C_{GL}]$$  \hspace{1cm} (11)

Accordingly, augmentation $Q_k$’s grade level is considered the maximum of the grade levels associated with the set of concepts $C(Q_k)$ present in the augmentation.

$$Grade(Q_k) = \max_{c_j \in C(Q_k)} Grade(c_j)$$  \hspace{1cm} (12)
Augmentations for school-level Textbooks

Fig. 7 A snapshot of the homepage of our user interface

$Q_{\text{final}}$, the final set of augmentations, was formed by filtering out the QA pairs $Q_k$ from $Q_{\text{rerank}}$ if its grade level was higher than $GL$.

$$Q_{\text{final}} = Q_{\text{rerank}} - \{ Q_k \mid \text{Grade}(Q_k) > GL \}$$ (13)

8 Developing a web-based interface

The present work proposed a web-based interface to achieve the annotation tasks and cater the video lectures and their augmentations to the students. Based on the type of users, this platform offers two different interfaces: annotator’s interface and student’s interface. Through login details of the homepage, users are redirected to an interface that suits their purpose. The homepage is shown in Fig. 7.

8.1 Annotator’s Interface

Each annotator is assigned a set of textbooks of specific subject and grade levels according to their expertise. Upon logging in, the annotators are redirected to either annotation tasks: identifying concept deficiencies or determining relevance of the recommended augmentations. For the former task, annotators are shown the textbook contents with hyperlinked concepts. A pop-down option will appear on selecting a concept, and the annotator can set a deficiency tag for the selected concept. After selecting deficiency tags for all the concepts of a textbook chapter, the annotator can submit the response. Figure 8 shows a snapshot of the annotator’s interface associated with the task of identifying concept deficiencies.

To determine relevance of the recommended augmentations, annotators are shown the textbook contents with concepts linked to their augmentations. The recommended augmentations will appear in the right panel on selecting a concept, and the annotator can set a relevance tag (relevant and irrelevant) for each augmentation. After selecting relevance for augmentations recommended for all the concepts, the annotator’s response can be submitted and stored. A snapshot of the annotator’s interface associated with this task is shown in Fig. 9.
Students are redirected to the students’ interface that offers a list of textbooks across subjects and grade levels from the homepage. Figure 10 shows this interface. On obtaining a specific textbook as an input, the system loads the textbook sections and extracts the concepts from the section. The input textbook contents appear on the interface chapter-wise, with concepts present in the chapter linked with the corresponding augmentations in the form of question-answer pairs. The screenshots of the students’ interface are shown in Fig. 11, which shows textbook contents of the 2nd chapter from the “Biology” textbook of grade level 11.

8.2 Student’s Interface

Students are redirected to the students’ interface that offers a list of textbooks across subjects and grade levels from the homepage. Figure 10 shows this interface.

On obtaining a specific textbook as an input, the system loads the textbook sections and extracts the concepts from the section. The input textbook contents appear on the interface chapter-wise, with concepts present in the chapter linked with the corresponding augmentations in the form of question-answer pairs. The screenshots of the students’ interface are shown in Fig. 11, which shows textbook contents of the 2nd chapter from the “Biology” textbook of grade “11.”
When a student clicks on a certain concept in the augmented textbook content, the linked augmentation in the form of relevant QA pair appears in the same window. A sample augmentation is shown in Fig. 12 for a concept “cyanobacteria”, selected from the 2nd chapter from “Biology” textbook of grade “11”.

The current work furnished the same set of augmentations for all students. However, students can choose between the concepts that need further elaborations.
The evaluation focused on two broad aspects of the proposed system: 1) automation to detect conceptual deficiencies and relevant augmentations and 2) effect of proposed cQA-based augmentations in the learning process. While the automation aspect is evaluated with respect to a set of carefully developed gold standard data, effect of cQAs in learning has been studied using interactions with students and teachers. With this focus, we have categorized the evaluation tasks as follows.

1. System Based Evaluation
   (a) Deficiency Diagnosis
   (b) Augmentation retrieval

2. Human Based Evaluation
   (a) Student Performance Based Evaluation of Augmentations (Quantitative)
   (b) Subject-Matter Expert Based Evaluation of Augmentations (Qualitative)
   (c) Student Based Evaluation of User Interface (Qualitative)

9.1 Corpus

To evaluate the proposed deficiency diagnosis model, the current work collected a corpus of textbooks as well as created a gold-standard data tagging the concepts present in it. Concerned QA archives were also collected from the concerned Stack Exchange sites to retrieve cQA-based augmentations.
9.1.1 Textbook data

The present work collected a corpus of Indian school-level textbooks issued by the National Council of Educational Research and Training (NCERT). The corpus, collected from 28 books of grade labels 6-12, covers seven subjects: Science, Mathematics, Chemistry, Physics, Biology, Geography, and Economics. This corpus consists of 68,264 concepts distributed across 8,062 sections, and 383 chapters.

9.1.2 Question-answer (QA) archive

A QA archive was collected from six Stack Exchange sites related to concerned textbook subjects: Mathematics, Chemistry, Physics, Biology, Economics, and Earth Science. Detailed statistics of the QA archive are mentioned in Table 2.

9.1.3 Gold-standard data for deficiency diagnosis

The gold standards were created by assigning 12 teachers who have been teaching the respective subjects in different schools for at least five years. We assigned the annotation task for six subjects (Physics+Science, Chemistry, Mathematics, Biology, Geography, and Economics) such that each textbook was assigned to 2 teachers. They were asked to annotate the concepts with their deficiency classes (D1 / D2 / D3 / D4). The inter-annotator agreement was measured using Cohen’s kappa (McHugh 2012) from the annotation matrix, presented in Table 3.
For classes D1, D2, D3, and D4, kappa values of 0.75, 0.72, 0.62, and 0.81 were achieved. While annotators agreed on D1, D2, and D4 classes to a satisfactory level, agreement for D3 is relatively low. The annotators were consulted on a subset of cases of disagreement in D3. However, the agreement did not improve significantly. This can be attributed to the fact that D3 is somewhat more subjective than the other classes as per the views of the annotators.

### 9.1.4 Gold-Standard Data for Augmentation Retrieval

The gold standards were created by assigning the same 12 teachers to annotate augmentations as relevant or irrelevant. We assigned the annotation task for six subjects (Physics+Science, Chemistry, Mathematics, Biology, geography, and Economics) such that each textbook is assigned to 2 teachers. The annotation matrix is presented in Table 4 while the inter-annotator agreement was measured using Cohen's kappa (McHugh 2012). For this annotation task, we achieved a satisfactory agreement with a kappa value of 0.71.

### 9.2 System based evaluation

The task of assessing performance of the constituent modules of our proposed textbook augmentation system raises following research questions.

1. SYS-RQ1: How effective is the deficiency diagnosis module?
2. SYS-RQ2: How effective is the augmentation retrieval module?

#### 9.2.1 SYS-RQ1 results

In the absence of a study on diagnosing concept-specific deficiencies, the current work used the section-specific variant presented by Agrawal et al. as baseline (Agrawal et al. 2012a). Considering the section-specific features mentioned in Section 5, this baseline system classified the deficiencies. The baseline was implemented using an SVM classifier similar to the proposed system. To evaluate the significance of the features, the dependence between the features were computed using the F-statistic (Guyon et al. 2002), chi square-statistic (Zhai et al. 2018), Bagged decision tree (Tran et al. 2017), and PCA (Song et al. 2010). If considered insignificant according to the above-mentioned statistics, a feature was not used while predicting the deficiency of a concept. This feature selection criterion was used for both baseline and proposed approaches. Choice of kernels and values for the parameters

| Annotator 1 | Annotator 2 |
|-------------|-------------|
| Relevant    | 13484       | 2249        |
| Irrelevant  | 2227        | 12453       |

Table 4 Inter-annotator Agreement matrix for augmentation retrieval
were determined similarly. While training and testing, we handled the issue of data imbalance by randomly selecting the same number of instances from each deficiency group (D1, D2, D3 and D4). Performance of the baseline and the proposed classification-based approaches were evaluated using standard metrics: Accuracy, Hamming loss, Precision score, Recall score, F1 score, Jaccard score, and Zero-one loss. Predictions for both the approaches were tested against the gold standard data, as shown in Tables 5 and 6. For some measures, a bigger value is better (such as accuracy, precision score, recall score, F1 score, and Jaccard score). On the other hand, a lesser value is better for two measures: Hamming loss and zero-one loss. These two categories of measures are denoted in these Tables with \(↑\) and \(↓\), respectively.

From both these tables, it is observed that the proposed approach (i) modeled all the deficiency classes better than the baseline, and (ii) the proposed approach
performed almost equally for all the subjects. Error cases in diagnosing deficiencies are primarily attributed to the extraction of irrelevant concepts. The concept extraction module may often return irrelevant concepts due to the unavailability of glossary terms. Consequently, its related, prerequisite, or dependent concepts are often noted missing in the concerned sections. Such concepts are often falsely predicted as “not deficient” due to a lack of significance in their context. A few cases are also reported due to poor recall in concept extraction. The number of extracted concepts gets lowered if one restricts the concept extraction module. Although the well-established entities are present in the textual discourse, they are not extracted as a concept. Consequently, one fails to locate the related, prerequisite, or dependent concepts for any concept. Due to a lack of precision in the concept extraction, the extracted concepts are often predicted deficient, although all its related concepts are present in the concerned section.

9.2.2 SYS-RQ2 Results

In the present study, we analyzed the performance of the augmentation retrieval module by determining the relevance of the recommended augmentations. We showed each augmentation to a teacher associated with teaching the concerned subject in different schools and asked to tag its binary relevance (relevant or irrelevant). Performance of the augmentation retrieval module is determined using six standard metrics: MAP (mean average precision), MRR (Mean reciprocal rank), RP (R-precision), P@1 (Precision at 1), P@5 (Precision at 5), and P@10 (Precision at 10) in three stages (Retrieval; Retrieval + Reranking; and Retrieval + Reranking + Filtering), as shown in Table 7.

The overall retrieval performance is satisfactory. We achieved consistently better performance by applying reranking and filtering steps over the initial retrieval approach. A few cases are reported where no augmentations have been retrieved as (i) relevant QA pairs were missing in the associated archive or (ii) relevant QA pairs were aligned to a higher grade level. On the other hand, the suggested augmentations suffered due to poor mapping between queries and QA pairs. Due to difficulty in representing deficiencies of a concept in natural languages, the present application forms a key-word based query using the concept and its context. Consisting of numerous concepts, often, intent of the query cannot be accurately mapped with relevant QA pairs. Consequently, irrelevant augmentations are suggested.
We replied upon the opinions of students and expert teachers to judge quality of the augmentations and usability of the user interface. These assessment tasks have formed following research questions:

1. **HUM-RQ1**: Do the augmented textbooks contribute to the learning gain?
2. **HUM-RQ2**: How satisfactory are the augmentations in the view of the teachers?
3. **HUM-RQ3**: How usable is the interface for presenting the augmentations?

### 9.3.1 HUM-RQ1 Results

To conduct a quantitative evaluation for the proposed augmentations, the current work experimented with 200 students enrolled in grade levels 9-12. The experiment comprised five major tasks. The experimental-flow is depicted in Fig. 13 to compare two types of study-materials: TB and ATB. TB is a normal textbook chapter, and ATB is the enhanced TB with the augmentations inserted at appropriate places.

The details of the tasks are mentioned below.

- **Grouping students (Task 1)**: The students were randomly divided into two homogeneous groups: Control group ($G_C$) and Experimental group ($G_E$) using simple randomization technique (Kim and Shin 2014). However, this division ensured equal number of students from the same grade level in both the groups.

- **Forming questionnaire (Task 2)**: Two sets of question papers, Q1 and Q2, each comprising ten questions from textbook chapter TB, were developed to check the basic conceptual and supplementary references among the students. Questions referring to the basic conceptual gap and supplementary references were popul-
lated in the equal distribution in both the questionnaires. These questionnaires are formed using teachers teaching the respective subjects in the concerned grade level in schools for at least 5 years. All these questions are within the scope of the curriculum and associated textbook. Content validity for these questionnaires were determined following (Lawshe 1975). According to Lawshe, the questions should be judged based on how useful they are in determining a given task and classified within ‘essential,’ ‘useful, but not essential,’ or ‘not necessary’ groups. If at least half the experts say that a question is ‘essential,’ the concerned question should have some content validity. We performed this test by asking 6 subject matter experts (teachers) for each question and calculated its content validity ratio \( CVR \) as:

\[
CVR = \frac{N_e - \frac{N}{2}}{\frac{N}{2}}
\]  

(14)

where \( N_e \) is the number of experts who tagged the concerned question as ‘essential’ and \( N \) is the total number of experts. Averaging the content validity ratio \( CVR \) for all the questions, content validity ratio \( CVR \) of a questionnaire is finally determined. The content validity ratio \( CVR \) value ranges from +1 to -1, while positive values indicate that at least half the experts rated the question as essential. For our questionnaires Q1 and Q2, the average content validity ratios are 0.49 and 0.56, respectively.

- **Pretest (Task 3):** Pretest with Q1 was conducted to judge the learning levels of the students from both groups before they study the associated materials. For each correct answer in the tests, 1 mark was awarded to the students.

- **Study Session (Task 4):** A study session was arranged on the same day pretest session had been arranged. In this session, students from \( G_E \) were offered with augmented textbook \( ATB \), whereas students from \( G_C \) were provided standard

| Table 8 Descriptive statistics of the student performance |  |
| --- | --- |
| Group | Test | N | Max | Min | Mean | SD |
| \( G_C \) | Pre-test | 100 | 4 | 1 | 2.11 | 0.82 |
| \( G_C \) | Post-test | 100 | 5 | 1 | 2.74 | 0.85 |
| \( G_E \) | Pre-test | 100 | 4 | 1 | 2.25 | 0.69 |
| \( G_E \) | Post-test | 100 | 5 | 2 | 3.43 | 0.89 |

| Table 9 Divergence between the distributions |  |
| --- | --- |
| Pair of distributions | Kullback-Leibler divergence |
| \( G_C \) pretest than \( G_E \) pretest | 0.02 |
| \( G_C \) posttest than \( G_C \) pretest | 0.29 |
| \( G_E \) posttest than \( G_E \) pretest | 0.37 |
| \( G_E \) posttest than \( G_C \) posttest | 0.1 |

- **Pretest (Task 3):** Pretest with Q1 was conducted to judge the learning levels of the students from both groups before they study the associated materials. For each correct answer in the tests, 1 mark was awarded to the students.

- **Study Session (Task 4):** A study session was arranged on the same day pretest session had been arranged. In this session, students from \( G_E \) were offered with augmented textbook \( ATB \), whereas students from \( G_C \) were provided standard

| Table 8 |  |
| --- | --- |
| Group | Test | N | Max | Min | Mean | SD |
| \( G_C \) | Pre-test | 100 | 4 | 1 | 2.11 | 0.82 |
| \( G_C \) | Post-test | 100 | 5 | 1 | 2.74 | 0.85 |
| \( G_E \) | Pre-test | 100 | 4 | 1 | 2.25 | 0.69 |
| \( G_E \) | Post-test | 100 | 5 | 2 | 3.43 | 0.89 |

| Table 9 |  |
| --- | --- |
| Pair of distributions | Kullback-Leibler divergence |
| \( G_C \) pretest than \( G_E \) pretest | 0.02 |
| \( G_C \) posttest than \( G_C \) pretest | 0.29 |
| \( G_E \) posttest than \( G_E \) pretest | 0.37 |
| \( G_E \) posttest than \( G_C \) posttest | 0.1 |
textbook content \( TB \). To nullify Hawthorne effect (Spano 2006), (i) no students were informed about their engagements, and (ii) the textbook chapters were selected such that the participating students had been able to understand the contents but had not studied the chapter earlier.

- **Post-test (Task 5):** Three days after the study session, a test was conducted using questionnaire Q2. Like the pretest, 1 mark was awarded to the students for each correct answer.

The descriptive statistics regarding the performance of the students are mentioned in Table 8 for four subjects: Science, Geography, Physics, and Biology.

To measure the quality of the suggested augmentations, the present work calculated the Kullback-Leibler divergence between the performance observed for the following pair of distributions. The divergence values are mentioned in Table 9.

Homogeneity of the control and experimental groups is established as no significant performance differences have been observed between the distributions \( G_C \) pretest and \( G_E \) pretest (divergence value of 0.02). Significance for learning materials (both \( TB \) and \( TB \)) is proven as the compared distributions are significantly different (divergence value of 0.29 and 0.37, respectively). However, the augmented textbook chapter \( ATB \) is found slightly more significant in developing better learning labels among the students (divergence value of 0.1).

The overall effect of the suggested augmentations is illustrated in Fig. 14 using boxplot representations.

Figures 15 and 16 presented the effect of augmentations in achieving better learning gains for different grade and subject domains.
Fig. 15  Students’ Performance Variance for Different Grade Levels

Fig. 16  Students’ Performance Variance for Different Subjects
From these diagrams, it is observed that: (i) for all the posttest scenarios, performance of $G_E$ surpassed the same of $G_C$ and (ii) performance of $G_E$ is significantly better for higher grade levels or associated subjects. The overall improvement in learning levels among the students from $G_E$ establishes the effectiveness of the proposed augmentations across the grade levels and subjects.

### 9.3.2 HUM-RQ2 Results

To evaluate the augmentations qualitatively, the current work assigned 12 teachers, teaching the respective subjects for at least five years in different schools. The survey is performed using a 5-point (score 1 refers to very poor and 5 to very good) Likert scale-based satisfaction questionnaire, consisting of 5 questions, as mentioned in Table 10.

The opinions from the teachers are mentioned in Fig. 17 for each of the survey questions. Analyzing the reliability of the questionnaire using Cronbach alpha (Cronbach and Shavelson 2004), the internal concordances were found to be within allowed ranges (0.78).

### Table 10 Questions from the Satisfaction Questionnaire

| Question | Description                                    |
|----------|------------------------------------------------|
| Question 1 | Rate the augmentations in developing basic concepts |
| Question 2 | Rate the augmentations in offering reference knowledge |
| Question 3 | Rate the augmentations when applied for practical subjects |
| Question 4 | Rate the augmentations when applied for theoretical subjects |
| Question 5 | Rate the augmentations to create interests among students |

![Fig. 17 Opinions of the subject-expert teachers on the suggested augmentations](image)
According to the qualitative evaluation, 45 out of 60 responses achieved a score of 3 or more for all the questions. This tells us that the augmentations achieved an overall satisfactory level. Although the augmentations offered relevant QA pairs to address the textbook deficiencies and additional references on related concepts, there are scopes to improve in terms of creating learning interests. Learning interests are commonly attributed to the student’s metacognition and situational factors (hsun Tsai et al. 2018). We can model these factors to retrieve augmentations that trigger learning interests among the students.

9.3.3 HUM-RQ3 Results

To assess the usability of our interface, 15 students were asked to submit their responses against the questions/statements from two standard questionnaires: Computer System Usability Questionnaire (Brooke 1996), and System Scale (Lewis 1995), respectively named as “SUS” and “IBM.”

- **SUS**: This questionnaire contains 10 questions. A response for each question can be given in a 5-point (0 to 4; where score 0 denotes to “completely disagree” and 4 to “completely agree”) Likert scale. With this range of values, the highest possible score is $4 \times 10 = 40$. This score is further converted to a 0 to 100 scale by multiplying with 2.5. The responses for each question are stated in Table 11 from 15 students (P1–P15 denote to the 15 students). Averaging the score for the responses obtained from 15 students for each question, we determined the usability score of our interface. Our designed interface achieved a usability score of 58.2 against the standard acceptable score of 54.4 Bangor et al. (2009). This score is considered valid considering the minimum sample size for the SUS test as at least 12 participants (Lewis and Sauro 2009). We analyzed the student responses to validate this questionnaire using Cronbach alpha (Cronbach and Shavelson 2004) and achieved a favorable internal concordance of 0.72. For two

| Participants | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 | P12 | P13 | P14 | P15 |
|--------------|----|----|----|----|----|----|----|----|----|-----|-----|-----|-----|-----|-----|
| Q1           | 3  | 3  | 4  | 2  | 3  | 4  | 3  | 4  | 2  | 4   | 3   | 4   | 3   | 3   | 3   |
| Q2           | 4  | 4  | 3  | 4  | 3  | 3  | 3  | 3  | 3  | 2   | 4   | 3   | 3   | 3   | 2   |
| Q3           | 2  | 2  | 3  | 2  | 4  | 3  | 3  | 3  | 3  | 3   | 3   | 4   | 4   | 3   | 4   |
| Q4           | 4  | 3  | 3  | 4  | 2  | 4  | 3  | 2  | 4  | 2   | 3   | 3   | 2   | 2   | 3   |
| Q5           | 2  | 1  | 3  | 2  | 3  | 3  | 2  | 3  | 3  | 3   | 2   | 2   | 3   | 1   | 2   |
| Q6           | 2  | 3  | 3  | 1  | 2  | 3  | 2  | 1  | 2  | 1   | 3   | 2   | 3   | 2   | 2   |
| Q7           | 4  | 2  | 2  | 3  | 3  | 2  | 4  | 2  | 3  | 3   | 3   | 3   | 2   | 2   | 1   |
| Q8           | 3  | 4  | 2  | 4  | 3  | 4  | 3  | 3  | 3  | 3   | 4   | 2   | 3   | 4   | 3   |
| Q9           | 3  | 4  | 2  | 3  | 4  | 3  | 3  | 2  | 4  | 3   | 4   | 2   | 4   | 3   | 3   |
| Q10          | 4  | 3  | 3  | 4  | 4  | 3  | 2  | 4  | 3  | 2   | 3   | 4   | 2   | 3   | 2   |
questions (e.g., questions numbered 5 and 6), the attained scores were noted to be comparatively lower. However, these questions address functional aspects of the interface rather than usability.

- **IBM**: Questionnaire IBM is formed with 19 questions, each responded using a 7-point (1 to 7; score 1 referring to “completely agree” and 7 to “completely disagree”) Likert scale. We presented the responses from 14 students for each question in Table 12, as one student did not respond to all the questions. The average score for 19 statements from 14 students was 1.86, with a standard deviation of around 0.7. This denotes usability of an overall satisfactory level. We have also analyzed the student responses to validate this questionnaire using Cronbach alpha (Cronbach and Shavelson 2004) and achieved a favorable internal concordance of 0.94. However, we noted a few statements for which the average score comes to 2.2+ with a standard deviation of 0.7+. This result tells us to improve the interface’s functionality, but a satisfactory level for usability.

### Table 12  Opinion of students on usability of the learning platform using questionnaire IBM

| Participants | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 | P12 | P13 | P14 |
|--------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| Q1           | 2  | 1  | 2  | 2  | 1  | 2  | 2  | 1  | 1  | 2  | 1  | 1  | 1  | 2  |
| Q2           | 1  | 1  | 2  | 1  | 2  | 2  | 1  | 1  | 1  | 1  | 2  | 2  | 1  | 2  |
| Q3           | 1  | 2  | 1  | 2  | 1  | 1  | 1  | 1  | 1  | 2  | 1  | 1  | 1  | 1  |
| Q4           | 2  | 1  | 2  | 2  | 3  | 1  | 3  | 1  | 1  | 1  | 1  | 1  | 2  | 1  |
| Q5           | 1  | 3  | 2  | 1  | 1  | 2  | 2  | 2  | 1  | 1  | 2  | 1  | 2  | 3  |
| Q6           | 3  | 2  | 1  | 2  | 3  | 1  | 1  | 1  | 2  | 1  | 2  | 3  | 2  | 1  |
| Q7           | 2  | 2  | 2  | 3  | 3  | 2  | 2  | 2  | 2  | 2  | 2  | 4  | 1  | 2  |
| Q8           | 1  | 2  | 3  | 3  | 3  | 2  | 3  | 3  | 2  | 4  | 2  | 3  | 2  | 3  |
| Q9           | 2  | 3  | 2  | 3  | 1  | 2  | 2  | 2  | 3  | 2  | 2  | 3  | 3  | 2  |
| Q10          | 1  | 2  | 2  | 2  | 2  | 2  | 2  | 2  | 2  | 1  | 3  | 2  | 3  | 1  |
| Q11          | 2  | 1  | 1  | 1  | 2  | 1  | 3  | 2  | 2  | 1  | 1  | 2  | 1  | 2  |
| Q12          | 3  | 3  | 2  | 2  | 1  | 2  | 1  | 2  | 3  | 2  | 1  | 2  | 1  | 2  |
| Q13          | 3  | 2  | 2  | 4  | 2  | 2  | 3  | 4  | 2  | 3  | 2  | 3  | 3  | 4  |
| Q14          | 1  | 1  | 3  | 1  | 2  | 1  | 2  | 1  | 1  | 3  | 2  | 1  | 2  |
| Q15          | 2  | 1  | 1  | 1  | 3  | 2  | 2  | 2  | 2  | 1  | 3  | 2  | 1  |
| Q16          | 2  | 3  | 2  | 1  | 1  | 2  | 2  | 1  | 1  | 3  | 2  | 2  | 1  | 2  |
| Q17          | 2  | 2  | 1  | 2  | 3  | 1  | 2  | 3  | 1  | 1  | 2  | 2  | 2  |
| Q18          | 1  | 1  | 2  | 2  | 1  | 2  | 3  | 1  | 1  | 2  | 2  | 1  | 3  |
| Q19          | 2  | 1  | 2  | 2  | 2  | 2  | 2  | 2  | 2  | 1  | 2  | 2  | 1  |

10 Conclusion

The present application offered a novel perspective for textbook augmentation by automatically aligning concepts in the learning materials with relevant QA pairs. To summarize, our proposed augmentations are considered helpful to an average
student as well as to a motivated student. We surveyed the existing studies to list the factors that affect the overall readability and conceptual deficiencies in the textbooks. A supervised machine-learning-based approach was adopted to diagnose these concept deficiencies. Our deficiency diagnosis model can be extended in developing related tasks (e.g., determining textbook quality, listing issues for designing course contents, or creating guidelines for writing textbooks). Keyword-based queries were formed by combining the concepts with their context and deficiency patterns. The queries were used to retrieve the QA pairs from the cQA archives. The relevant QA pairs were then recommended to the students. We presented system and human-based evaluation strategies to analyze the performance of the proposed augmentation system, the overall quality of the augmentations, and the usability of the interface.

Despite our proposed system performed to a satisfactory level, there are certain limitations. The concept extraction and deficiency diagnosis models depend on the link structure of the Wikipedia knowledge base. The absence of a Wikipedia title corresponding to a candidate concept leads to inappropriate pruning of the concept. The deficiency diagnosis model made use of aspects like related concepts, prerequisite concepts, and others that had been extracted by applying some heuristics over Wikipedia link structure. These structures are induced by the concepts (and aligned Wikipedia title) that are being probed. Incorrect concept to Wikipedia title mapping that may arise due to ambiguity may lead to erroneous link structure and consequently extraction of incorrect concept aspects. The retrieval performance, on the other hand, highly depends on the size of associated QA archives and their comprehension level. Consequently, a small number of augmentations have been recommended for lower grade levels and subjects having smaller QA archives. It is to be noted that the augmentations are student agnostic. We believe that the mentioned limitations may lead to research studies that aim to automate the process of improving learning materials.

**Acknowledgements** The author is grateful to the students and teachers from Mount Litera Zee School, Nalanda and Jion Marshal English Medium School, Kharagpur, for their participation in the meticulous annotation and assessment tasks.

**Funding** Not applicable

**Data Availability Statement** Data used in the current work is shared using Google drive. The detail is provided at a Github repository present at: https://github.com/KrishnenduGhosh/TB

**Code Availability** Codes for implementing the current work is provided at as a Github repository. The link is: https://github.com/KrishnenduGhosh/TB

**Declarations**

**Conflicts of interest** Not applicable
References

Adesope, O. O., Cavagnetto, A., Hunsu, N. J., Anguiano, C., & Lloyd, J. (2017). Comparative effects of computer-based concept maps, refutational texts, and expository texts on science learning. *Journal of Educational Computing Research, 55*(1), 46–69.

Agrawal, R., Chakraborty, S., Gollapudi, S., Kannan, A., Kenthapadi, K. (2012a). Empowering authors to diagnose comprehension burden in textbooks. In: Proceedings of the 18th ACM SIG-KDD international conference on knowledge discovery and data mining (pp. 967–975). ACM

Agrawal, R., Christoforaki, M., Gollapudi, S., Kannan, A., Kenthapadi, K., Swaminathan, A. (2014). Mining videos from the web for electronic textbooks. In: International conference on formal concept analysis (pp. 219–234). Springer

Agrawal, R., Gollapudi, S., Kannan, A., Kenthapadi, K. (2011a). Identifying enrichment candidates in textbooks. In: Proceedings of the 20th international conference companion on World wide web (pp 483–492). ACM

Agrawal, R., Gollapudi, S., Kannan, A., Kenthapadi, K. (2011b). Enriching textbooks with images. In: Proceedings of the 20th ACM international conference on Information and knowledge management (pp. 1847–1856). ACM

Agrawal, R., Gollapudi, S., Kannan, A., Kenthapadi, K. (2012). Data mining for improving textbooks. *ACM SIGKDD Explorations Newsletter, 13*(2), 7–19.

Al-Qatawneh, S., Alsalhi, N., Al Rawashdeh, A., Ismail, T., & Aljarrah, K. (2019). To e-textbook or not to e-textbook? a quantitative analysis of the extent of the use of e-textbooks at ajman university from students’ perspectives. *Education and Information Technologies, 24*(5), 2997–3019.

Alsalem, M. A. (2018). Exploring metacognitive strategies utilizing digital books: Enhancing reading comprehension among deaf and hard of hearing students in saudi arabian higher education settings. *Journal of Educational Computing Research, 56*(5), 645–674.

Anderson, T. H., Armbruster, B. B., et al. (1984). Content area textbooks. Learning to read in American Schools: Basal Readers and Content Texts, 193–226

Aritajati, C. & Narayanan, N. H. (2013). Facilitating students’ collaboration and learning in a question and answer system. In: Proceedings of the 2013 conference on Computer supported cooperative work companion (pp. 101–106)

Asheim, L. (1958). Readability: An appraisal of research and application (book review). *College & Research Libraries, 19*(5), 422–423.

Augustijn, E.-W., Lemmens, R., Verkroost, M.-J., Ronzhin, S., & Walsh, N. (2018). The living textbook: Towards a new way of teaching geo-science. Lund, Sweden: Agile.

Babinec, P. & Srba, I. (2017). Education-specific tag recommendation in cqa systems. In: Adjunct publication of the 25th conference on user modeling, adaptation and personalization (pp. 281–286). ACM

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The berkeley framenet project. In: 36th Annual meeting of the association for computational linguistics and 17th international conference on computational linguistics, vol .1 (pp. 86–90)

Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of Usability Studies, 4*(3), 114–123.

Bennani-Smires, K., Musat, C., Hossmann, A., Baeriswyl, M., & Jaggi, M. (2018). Simple unsupervised keyphrase extraction using sentence embeddings. arXiv:1801.04470, (pp. 221–229)

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*(1–7), 107–117.

Brooke, J. (1996). SUS-A quick and dirty usability scale. Usability evaluation in industry. CRC Press. ISBN: 9780748404605

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., Jatowt, A. (2018). A text feature based automatic keyword extraction method for single documents. In: European conference on information retrieval (pp. 684–691). Springer

Carreira-Perpinan, M. A. & Hinton, G. E. (2005). On contrastive divergence learning. In: Aistats, vol. 10, (pp. 33–40). Citeseer.

Chambliss, M. & Calfee, R. (1998). Textbooks for learning: Nurturing children’s minds. Blackwell Publishing.
Chang, C.-C. & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chen, D. & Manning, C. (2014). A fast and accurate dependency parser using neural networks. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 740–750)

Chen, Y., Gao, L., Shi, S., Du, X., Wen, J.-R. (2014). Improving context and category matching for entity search. In: Proceedings of the twenty-eighth AAAI conference on artificial intelligence, July 27 -31, 2014, Qu'ebeec City, Qu'ebeec, Canada (pp. 16–22)

Choi, E., Borkowski, M., Zakoian, J., Sagan, K., Scholla, K., Ponti, C., et al. (2015). Utilizing content moderators to investigate critical factors for assessing the quality of answers on brainly, social learning g&q&a platform for students: a pilot study. Proceedings of the Association for Information Science and Technology, 52(1), 1–4.

Clark, R. C., Nguyen, F., Sweller, J. (2011). Efficiency in learning: Evidence-based guidelines to manage cognitive load. John Wiley & Sons.

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. Journal of Applied Psychology, 60(2), 283.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. Educational and Psychological Measurement, 64(3), 391–418.

Crossley, M., & Murby, M. (1994). Textbook provision and the quality of the school curriculum in developing countries: Issues and policy options. Comparative Education, 30(2), 99–114.

Daiber, J., Jakob, M., Hokamp, C., Mendes, P. N. (2013). Improving efficiency and accuracy in multilingual entity extraction. In: Proceedings of the 9th international conference on semantic systems (i-semantics).

Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. Educational Research Bulletin, 37–54

DuBay, W. H. (2004). The principles of readability. Online Submission

El-Beltagy, S. R., & Rafea, A. (2009). Kp-miner: A keyphrase extraction system for english and arabic documents. Information Systems, 34(1), 132–144.

Ewais, A., & Troyer, O. D. (2019). A usability and acceptance evaluation of the use of augmented reality for learning atoms and molecules reaction by primary school female students in palestine. Journal of Educational Computing Research, 57(7), 1643–1670.

Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology, 32(3), 221.

Garcia-Gonzalez, H., Gayo, J. E. L., Paule-Ruiz, M. (2016). Enhancing e-learning content by using semantic web technologies. IEEE Transactions on Learning Technologies

Gelderblom, H., Matthee, M., Hattingh, M., & Weilbach, L. (2019). High school learners’ continuance intention to use electronic textbooks: a usability study. Education and Information Technologies, 24(2), 1753–1776.

Ghosh, K., Bhowmick, P. K., Goyal, P. (2017). Using re-ranking to boost deep learning based community question retrieval. In: Proceedings of the International Conference on Web Intelligence, WI ’17, pages 807–814, New York, NY, USA. ACM

Ghosh, K., Nangi, S. R., Kanchugantla, Y., Rayapati, P. G., Bhowmick, P. K., Goyal, P. (2021). Augmenting video lectures: Identifying off-topic concepts and linking to relevant video lecture segments. International Journal of Artificial Intelligence in Education, 1–31

Gray, W. S. & Leary, B. E. (1935). What makes a book readable.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46(1–3), 389–422.

Hewitt, P. G. (2002). Conceptual physics. Pearson Education

hsun Tsai, T., hung Lin, C., chao Hong, J., & hsin Tai, K. (2018). The effects of metacognition on online learning interest and continuance to learn with moocs. Computers & Education, 121, 18–29.

Kempe, A.-L., & Grönlund, Å. (2019). Collaborative digital textbooks-a comparison of five different designs shaping teaching and learning. Education and Information Technologies, 24(5), 2909–2941.

Kieras, D.E. & Dechert, C. (1985). Rules for comprehensible technical prose: A survey of the psycholinguistic literature. Technical report, University of Michigan

Kim, J., & Shin, W. (2014). How to do random allocation (randomization). Clinics in Orthopedic Surgery, 6(1), 103–109.

Kingsbury, I. (2021). Online learning: How do brick and mortar schools stack up to virtual schools? Education and Information Technologies, pp 1–22

Kingsbury, I. (2021). Online learning: How do brick and mortar schools stack up to virtual schools? Education and Information Technologies, pp 1–22

Kingsbury, I. (2021). Online learning: How do brick and mortar schools stack up to virtual schools? Education and Information Technologies, pp 1–22

Springer
Klippel, A., Zhao, J., Jackson, K. L., La Femina, P., Stubbs, C., Wetzel, R., et al. (2019). Transforming earth science education through immersive experiences: Delivering on a long held promise. *Journal of Educational Computing Research, 57*(7), 1745–1771.

Koć-Januchta, M. M., Schönborn, K. J., Tibell, L. A., Chaudhri, V. K., & Heller, H. C. (2020). Engaging with biology by asking questions: Investigating students’ interaction and learning with an artificial intelligence-enriched textbook. *Journal of Educational Computing Research, 58*(6), 1190–1224.

Kopciwicz, L., & Bougiasa, H. (2021). Understanding emergent teaching and learning practices: iPad integration in Polish school. *Education and Information Technologies, 26*(3), 2897–2918.

Kumar, S. & Chauhan, A. (2019). Enriching textbooks by question-answers using cqa. In: TENCON 2019-2019 IEEE Region 10 Conference (TENCON) (pp. 707–714). IEEE

Kurland, O. & Lee, L. (2005). Pagerank without hyperlinks: Structural re-ranking using links induced by language models. In: Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR ’05, pp 306–313, New York, NY, USA. ACM

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 563–575.

Le, L. T., Shah, C., Choi, E. (2016). Evaluating the quality of educational answers in community question-answering. In: Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries (pp. 129–138). ACM

Leonard, S., Stroud, M.J., Shaw, R.J. (2021). Highlighting and taking notes are equally ineffective when reading paper or e-text. *Education and Information Technologies, 1–13*

Lewis, J. R. & Sauro, J. (2009). The factor structure of the system usability scale. In: International conference on human centered design, pp. 94–103. Springer.

Lewis, J. R. (1995). Ibm computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction, 57–78*

Lockheed, M. E., & Hanushek, E. (1988). Improving educational efficiency in developing countries: what do we know? *Compare, 18*(1), 21–38.

Macina, J., Srba, I., Williams, J. J., Bielikova, M. (2017). Educational question routing in online student communities. In: Proceedings of the eleventh ACM conference on recommender systems (pp. 47–55). ACM

Manning, C., Raghavan, P., & Schütze, H. (2010). *Introduction to information retrieval.* Natural Language Engineering, 16(1), 100–103.

McLaughlin, G. H. (1969). Smog grading—a new readability formula. *Journal of Reading, 12*(8), 639–646.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica: Biochemia Medica, 22*(3), 276–282.

Medelyan, O., Frank, E., Witten, I.H. (2009). Human-competitive tagging using automatic keyphrase extraction. In: Proceedings of the 2009 conference on empirical methods in natural language processing (pp. 1318–1327)

Mershad, K., Damaj, A., Wakim, P., Hamieh, A. (2019). Learnsmart: A framework for integrating internet of things functionalities in learning management systems. *Education and Information Technologies, 1–34*

Mihalcea, R. & Tarau, P. (2004). Textrank: Bringing order into text. In: Proceedings of the 2004 conference on empirical methods in natural language processing (pp. 404–411)

Mitra, S., & Gupta, S. (2020). Mobile learning under personal cloud with a virtualization framework for outcome based education. *Education and Information Technologies, 25*(3), 2129–2156.

Mohammad, R. & Kumari, R. (2007). Effective use of textbooks: A neglected aspect of education in Pakistan. *Journal of Education for International Development, 3*(1):1–12. 11/122.1212x

Mulvany, N. C. (2009). *Indexing books.* University of Chicago Press.

Nguyen, T.D. & Kan, M.-Y. (2007). Keyphrase extraction in scientific publications. In: International conference on asian digital libraries (pp. 317–326). Springer

Nipa, T. J., & Kermanshachi, S. (2020). Assessment of open educational resources (oer) developed in interactive learning environments. *Education and Information Technologies, 25*(4), 2521–2547.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*(1), 1–4.

Papagiannopoulou, E., & Tsoumakas, G. (2020). A review of keyphrase extraction. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(2), e1339.

Pawlowski, J., Barker, K., Okamoto, T. (2007). Quality research for learning, education, and training. *Reading & Writing Quarterly, 10*(2)

Purgina, M., Mozgovoy, M., & Blake, J. (2020). Wordbricks: Mobile technology and visual grammar formalism for gamification of natural language grammar acquisition. *Journal of Educational Computing Research, 58*(1), 126–159.
Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory, 1*, 1–20.

Schoevers, E.M., Leseman, P.P., Kroesbergen, E.H. (2019). Enriching mathematics education with visual arts: Effects on elementary school students’ ability in geometry and visual arts. *International Journal of Science and Mathematics Education, 1–22*

Schuyten, G., Dekeyser, H., & Goeminne, K. (1999). Towards an electronic independent learning environment for statistics in higher education. *Education and Information Technologies, 4(4), 409–424.*

Singh, M., Adebayo, S.O., Saini, M., Singh, J. (2021). Indian government e-learning initiatives in response to covid-19 crisis: A case study on online learning in indian higher education system. *Education and Information Technologies, 1–39*

Singh, G.K., Kumar, V., Bhat, S., Pedanekar, N. (2015). Automatically augmenting learning material with practical questions to increase its relevance. In: *Frontiers in Education Conference (FIE), 2015 IEEE* (pp. 1–7). IEEE

Song, F., Guo, Z., Mei, D. (2010). Feature selection using principal component analysis. In: *2010 International conference on system science, engineering design and manufacturing informatization, vol. 1, (pp. 27–30)*

Spano, R. (2006). Observer behavior as a potential source of reactivity: Describing and quantifying observer effects in a large-scale observational study of police. *Sociological Methods & Research, 34(4), 521–553.*

Srba, I. & Bielikova, M. (2015). Askalot: community question answering as a means for knowledge sharing in an educational organization. In: *Proceedings of the 18th ACM conference companion on computer supported cooperative work & social computing (pp. 179–182)*

Srba, I., Savic, M., Bielikova, M., Ivanovic, M., & Pautasso, C. (2019). Employing community question answering for online discussions in university courses: Students’ perspective. *Computers & Education, 135, 75–90.*

Toutanova, K., Klein, D., Manning, C.D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In: *Proceedings of the 2003 conference of the north american chapter of the association for computational linguistics on human language technology-volume 1, (pp. 173–180).* Association for Computational Linguistics.

Tran, C.T., Zhang, M., Andreae, P., Xue, B. (2017). Bagging and feature selection for classification with incomplete data. In: *European conference on the applications of evolutionary computation (pp. 471–486).* Springer

Tyson-Bernstein, H. (1988). *A Conspiracy of Good Intentions.* ERIC: America’s Textbook Fiasco.

Uluyol, C., & Agca, R. K. (2012). Integrating mobile multimedia into textbooks: 2d barcodes. *Computers & Education, 59(4), 1192–1198.*

Wan, X. & Xiao, J. (2008). Single document keyphrase extraction using neighborhood knowledge. In: *AAAI, vol. 8, (pp. 855–860)*

Wang, M., Chau, H., Thaker, K., Brusilovsky, P., He, D. (2021). *Knowledge annotation for intelligent textbooks (pp. 1–22). Knowledge and Learning: Technology.*

Wang, X., & Xing, W. (2019). Understanding elementary students’ use of digital textbooks on mobile devices: A structural equation modeling approach. *Journal of Educational Computing Research, 57(3), 755–776.*

Weng, C., Otanga, S., Christiano, S. M., & Chu, R.J.-C. (2020). Enhancing students’ biology learning by using augmented reality as a learning supplement. *Journal of Educational Computing Research, 58(4), 747–770.*

Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C., and Nevill-Manning, C.G. (2005). Kea: Practical automated keyphrase extraction. In: *Design and usability of digital libraries: Case studies in the asia pacific (pp. 129–152).* IGI global.

Woodward, A., Elliot, D.L., Nagel, K.C. (2013). *Textbooks in school and society: An annotated bibliography & guide to research, vol. 6.* Routledge.

Zhai, Y., Song, W., Liu, X., Liu, L., Zhao, X. (2018). A chi-square statistics based feature selection method in text classification. In: *2018 IEEE 9th International conference on software engineering and service science (ICSESS), (pp. 160–163). IEEE*

Zhou, G., Liu, Y., Liu, F., Zeng, D., Zhao, J. (2013). Improving question retrieval in community question answering using world knowledge. In: *Proceedings of the twenty-third international joint conference on artificial intelligence, IJCAI ’13 (pp. 2239–2245).* AAAI Press