Enriching Textbooks Through Data Mining

Rakesh Agrawal
Microsoft Research
Mountain View, CA, USA
rakesha@microsoft.com

Sreenivas Gollapudi
Microsoft Research
Mountain View, CA, USA
sreenig@microsoft.com

Krishnaram Kenthapadi
Microsoft Research
Mountain View, CA, USA
krisken@microsoft.com

Nitish Srivastava*
Indian Institute of Technology
Kanpur, India
nitishs@cse.iitk.ac.in

Raja Velu*
Syracuse University
Syracuse, NY, USA
rpvelu@syr.edu

ABSTRACT

Textbooks play an important role in any educational system. Unfortunately, many textbooks produced in developing countries are not written well and they often lack adequate coverage of important concepts. We propose a technological solution to address this problem based on enriching textbooks with authoritative web content. We augment textbooks at the section level for key concepts discussed in the section. We use ideas from data mining for identifying the concepts that need augmentation as well as to determine the links to the authoritative content that should be used for augmentation. Our evaluation, employing textbooks from India, shows that we are able to enrich textbooks on different subjects and across different grades with high quality augmentations using automated techniques.

1. INTRODUCTION

Education has long been acknowledged as one of the linchpins to improve the lives of the economically disadvantaged [1]. It can be the primary tool for improving people's abilities to be productive members of society. Many studies show a strong relationship between educational quality and wages. For example in [14], wages were found to be between 10 and 22 percent higher for each standard deviation increase in test results.

The problem of improving the quality of education is multifaceted and complex [6, 29, 31]. Bill Gates, in his 2007 Harvard Commencement address, put forward a framework for attacking complex problems [4]. We applied this framework to the problem of improving the quality of education in developing countries as shown in Figure 1 and concluded to concentrate our research on developing technological approaches for improving the quality of educational material. This paper reports on our progress and results obtained.

*Work done while visiting Microsoft Research.

ACM DEV'10, December 17–18, 2010, London, United Kingdom.

Framework	Application to Education
1. Define goal	1. Quality education to all
2. Find the highest leverage approach	2. New pedagogy
3. Discover the ideal technology for that approach	3. Individualized learning with teacher as discussant
4. In the meantime, make the smartest application of the technology on-hand.	4. Improve the quality of educational material

Figure 1: Framework for attacking complex problems, applied to improving education

1.1 Enriching Textbooks

Educational material comprises of a broad range of education related resources that includes textbooks, instructional guides, workbooks, practice exercises, activities, tests, and supplementary readers. We focus on textbooks as they are the primary vehicles for delivering content knowledge to students. Textbooks are also indispensable for fostering teacher learning and constitute a key component of ongoing professional development of the teachers [11, 27]. Multiple studies during last twenty five years have highlighted the positive impact of relevant, good-quality textbooks on student achievement. Particularly in developing countries, evidence suggests that textbooks are one of the most cost-effective means of positively impacting educational quality [7, 9, 10, 15, 16, 21, 22].

Unfortunately, many textbooks in developing countries suffer from the lack of clarity of language as well as the inadequacy of information provided in the textbook. Here is an example from a grade V Science textbook used in Baluchistan (cf. [25]):

The concept of 'lever' was defined as a "strong rod or stick on which force is applied on its one end and can be rotated through some support and work is done on the other end".

It was difficult for the teacher to see how this definition of 'lever' could qualify it as a machine when force was being exerted on one end and work was also done on the other end; the notion of 'input' and 'output' did not get clearly communicated. The teacher, therefore, asked the students to memorize the definition without helping them understand the scientific notion [25].

Because of cost considerations, textbooks are compressed into fewer pages resulting in poor exposition of subject matter. We quote from a critique of a grade IX Indian History textbook [23]:

The whole (medieval) period has been presented as a dull and dry history of dynasties, cluttered with the names and military conquests of kings, followed by brief acknowledgements of "social and cultural life", "art and architecture", "revenue administration", and so on. The entire Mughal period (1526-1707) is disposed of in six pages.

1.2 Scope of the paper

We investigate the techniques for enriching the content of textbooks. Our overall approach consists of the following steps:

1. Identify key concepts in a section of a chapter.
2. Find authoritative material on the Web for those concepts.
3. Add suitable number of links to the web content in the section.

The paper describes the algorithms used in our implementation and the results of our experimental evaluation of the approach. For sourcing supplemental material, we use Wikipedia.

We present the technology core for identifying links with which a unit of a textbook must be augmented, but do not discuss the details of the mechanisms for integrating those links into the textbook and how the augmented material is made available to students and teachers. Issues such as implications for royalty sharing and intellectual property rights are outside the scope of the paper. It has been pointed out that learning outcomes depend not only on the availability of textbooks, but also on how they are used by the teachers and how effectively have they been integrated with other interventions [11, 12, 24, 26]. While such deployment issues are critically important, they are beyond the scope of this work.

As observed in [3, 20], technology alone cannot solve complex societal problems such as providing universal, high-quality education; technological solutions have to dove-tail with the political imperatives, traditions and norms, individual and societal motivational factors, and the priority of other initiatives. However, by advancing technology options, we increase the policy choices and improve the overall quality of the solution. Therein lies the driving motivation for this work.

1.3 Paper Layout

The rest of the paper is organized as follows. We first discuss related work in Section 2. We then describe how we use data mining to enrich textbooks in Section 3. The experimental evaluation of our implementation is presented in Section 4. We conclude with a summary and directions for future work in Section 5.

2. RELATED WORK

IBM carried out an extensive study of how information technology can be used to improve India's education system [3]. The report identifies the deficiencies in textbooks and the lack of teaching support material as the major problems to which information technology can be brought to bear. The report argues for creating a nation-wide education network to harness the collective efforts of educators, parents, and students to monitor and collaboratively enhance the quality of educational material. The focus of our work is on enriching the deficient sections of textbooks with authoritative content mined from the web. Thus, the two efforts can be viewed as complementary approaches to improving the quality of textbooks.

A website that aims to enrich the learning experience with books is Notemonk.com. This website allows students to download textbooks, ask questions on a topic in a book, and annotate them for quick reference. Participants are also invited to contribute videos that can be subsequently viewed by other students. One issue we noted was that many of the videos had at best tangential relevance to the books and their quality varied dramatically. In our view, it is critical that the supplementary material provided is authoritative, must have high contextual relevance, and is linked as close to where the concept needing elaboration is presented in the text.

Related work to improve the quality of educational material includes efforts by several institutions to make the videos of the course lectures available through Internet. EducationPortal.com provides links to free video lectures from several prestigious universities. However, it is left to the initiative of individual students to find the relevant lectures to supplement their knowledge. Other issues that can hamper the utility of these otherwise extremely useful resources include language, pronunciation and accent of the speakers as well as difficulty to relate to examples and illustrations due to socio-economic and cultural differences [2]. A noteworthy effort that addresses some of these concerns is the Digital StudyHall (dsh.cs.washington.edu). They digitally record live classes by the best grassroots teachers, transmit them by DVDs through the postal system, collect them in a large distributed database, and distribute them on DVDs to poor rural and slum schools.

Many websites today enrich their web pages by showing definitions/auxiliary information when the reader hovers on keywords in a page. These augmentations are usually done by human editors. Some websites automatically link keywords to other web pages. Very often though these are advertisements and the keywords that are linked are determined by a bidding system. The linked pages frequently have very little to do with the content of the source page. Recent work, particularly in the context of contextual advertising, also extracts key phrases from a web page [28, 30, 34]. They employ learning algorithms that make use of features based on word frequencies, HTML meta data and query and clicks logs to detect the most important phrases. Such features are not available in our application.

3. METHODOLOGY

Our methodology for enriching textbooks consists of determining important concepts discussed in a section of a book, finding authoritative material on the Web for these concepts, and enhancing the section with suitable number of links to authoritative content.

3.1 Determining Key Concepts

Finding key concepts in a text is a challenging problem. The present approaches primarily involve detection of the key phrases based on rules (grammar) or statistical and learning methods. In the former, the structural properties of phrases form the basis for the rule generation. In the latter, the importance of a phrase is computed based on statistical properties (e.g., relative frequency, document frequency) of

Algorithm 1 DETERMINEKEYCONCEPTS

Input: A corpus of textbooks divided into sections; Grammatical pattern (regex) for detecting terminological noun phrases, R; Pruning parameter set, Θ.
Output: The set of key concept phrases for each section.

1: Tag every sentence in the book corpus using a part-of-speech tagger. (§3.1.1)
2: **for** each section i **do**
3: Compute the set C_i of noun phrases that maximally match the expression R. (§3.1.2)
4: Use a lexical database to identify phrases in C_i with inconsistent tagging, correct them if possible but otherwise drop the phrases. (§3.1.3)
5: Compute the probability scores of phrases in C_i using a web N-gram Service. (§3.1.4)
6: **end for**
7: Prune based on the score distribution of noun phrases over all sections and the pruning parameter set Θ to obtain the set of key concept phrases for each section. (§3.1.5)

the phrase [17].

After studying several textbooks, we devised the following approach for identifying key concepts in a unit of a book. *Concepts in our system correspond to terminological noun phrases.* We first form a candidate set of concepts using linguistic patterns, with the help of a part-of-speech tagger. We used two of the linguistic patterns proposed in [18] that have been used widely in the NLP community. We supplemented this set by a third pattern based on our inspection of the key concepts we identified by studying books on different subjects. We then exploit complementary signals from a different source, namely, a lexical database, to correct errors made by the part-of-speech tagger. Next we eliminate both malformed phrases and very common phrases, based on the probabilities of occurrences of these phrases on the Web. The reason for eliminating common phrases is that they would be already well understood and would not benefit as much from web augmentation.

Our implementation employs Stanford POS Tagger [32] for part-of-speech tagging, WordNet [8] as the lexical database, and Microsoft Web N-gram Service [33] to aid pruning of malformed and common phrases. Our methodology, however, is oblivious to the specific tools, though the performance of the system is dependent on them.

We state these steps in Algorithm 1 and discuss each step in detail below.

3.1.1 Part-of-speech Tagging

We tag every sentence in the book corpus using Stanford POS Tagger. The tagger assigns a unique part-of-speech to each word in a sentence, by processing the entire sentence. It predicts the part-of-speech even for an unknown word (such as a proper noun) by exploiting the context of the word in a sentence. Our corpus may contain poorly formed sentences, due to pdf parsing issues as well as presence of text extracted from tables, mathematical equations, and other non-grammatical structures. For such sentences, the assigned part-of-speech tags may be incorrect.

3.1.2 Detecting Terminological Noun Phrases

We next form a candidate set of concepts by determining the terminological noun phrases present in the text. The concepts of interest in our application typically consist of noun phrases containing adjectives, nouns, and sometimes prepositions. It is rare for concepts to contain other parts of speech such as verbs, adverbs, or conjunctions.

We consider three grammatical patterns (P_1, P_2, and P_3) for determining terminological noun phrases. The first two of these are from [18] and the third is the one we added.

We can express the three patterns using regular expressions as:

$$P_1 = C^*N$$

$$P_2 = (C^*NP)^?(C^*N)$$

$$P_3 = A^*N^+$$

where N refers to a noun, P a preposition, A an adjective, and $C = A|N$. P_1 corresponds to a sequence of zero or more adjectives or nouns, ending with a noun, while P_2 is a relaxation of P_1 that also permits two such patterns separated by a preposition. Examples of the former include "cumulative distribution function", "fiscal policy", and "electromagnetic radiation". Examples of the latter include "degrees of freedom" and "Kingdom of Asoka". P_3 corresponds to a sequence of zero or more adjectives, followed by one or more nouns. This pattern is a restrictive version of P_1, where an adjective occurring between two nouns is not allowed. The motivation for this pattern stems from sentences such as the following: "The experiment with Swadeshi gave Mahatma Gandhi important ideas about using cloth as a symbolic weapon against British rule". As a result of allowing arbitrary order of adjectives and nouns, "Mahatma Gandhi important ideas" is detected as a terminological noun phrase by pattern P_1. On the other hand, pattern P_3 would result in the better phrases, "Mahatma Gandhi" and "important ideas".

Note that our candidate concepts comprise of maximal pattern matches. Thus, we will not have "distribution function" as a candidate in presence of "cumulative distribution function". The intuition again is that it is better to augment more specific concepts than general concepts. Similar strategy was used in [19].

3.1.3 Correcting Errors using WordNet

Stanford POS Tagger can make errors on poorly formed sentences or on sentences containing unknown words. We experimented with using WordNet to detect these errors and correct them. WordNet is a large lexical database that groups words into sets of cognitive synonyms called synsets, each expressing a distinct concept. We use WordNet to determine possible parts of speech (noun, adjective, verb, adverb) for words in its knowledge base. However, WordNet would fail to recognize words absent in its database. WordNet being a hand curated system should have better accuracy than an automated parsing tool, but lower coverage. We therefore use WordNet as a validation and error-correcting tool.

We check whether the parts of speech assigned by Stanford POS Tagger is consistent with those provided by WordNet. We say that disagreement occurs for a phrase if for some word w in the phrase, (a) WordNet recognizes w and returns one or more part-of-speech tags *and* (b) the part-of-speech

tag assigned by Stanford POS Tagger is *not* one among the part-of-speech tags assigned by WordNet. For example, for the phrase "steatite micro beads", Stanford POS Tagger assignment is `<Adjective><Noun><Noun>` whereas WordNet assignment is `<Noun><Adjective><Noun>`. In such cases, we flag POS Tagger assignment as potentially incorrect, and correct the potential error using WordNet assignment if it can be uniquely corrected. In the above example, the assignment will be modified to `<Noun><Adjective><Noun>`. For the phrase "control measures", POS Tagger assignment is `<Adjective><Noun>` and WordNet assignment is `<Noun|Verb><Noun|Verb>`, resulting in disagreement, but we do not correct since there is no unique way to do so. For phrases with disagreements, we retain the phrase as part of the candidate set only if we can correct using WordNet assignment and if the corrected assignment still satisfies the grammatical pattern; in all other cases, we drop the phrase from the set.

3.1.4 Probability Scores using Web N-gram Service

The set of candidate phrases generated in the previous steps is likely to contain a number of common knowledge phrases as well as some malformed or unimportant long phrases. As our goal is to identify only the concepts that require further elaboration, we would like to prune such phrases.

For each phrase in the candidate set, we first obtain the probability of occurrence of the phrase on the Web using the Microsoft Web N-gram Service. We use this probability as a proxy for whether the phrase is part of common knowledge, since a common knowledge phrase is likely to have a significant presence on the Web. Similarly this probability can also indicate whether the phrase is malformed, as such phrases are less likely to occur on the Web. Thus, after obtaining the probability scores for each phrase, we compute the score distribution across phrases over the entire corpus, and prune based on this distribution to remove undesirable phrases.

Microsoft Web N-gram Service provides probability of occurrence of a phrase over three corpora: body of web pages, titles of pages, and anchor texts for web pages. We use the anchor corpus, since compared to title or body of a web page, anchor text provided a stronger signal. Anchor text can be viewed as representing how other web authors succinctly represent the target page.

3.1.5 Pruning using Statistical Properties of Score Distribution

Given the distribution D of N-gram log probability scores of candidate phrases across a corpus, we compute certain parametrized statistical boundaries. Let Q_1 denote the first quartile, that is, Q_1 satisfies $Pr_{x \in D}(x \leq Q_1) = 0.25$. Similarly let Q_3 denote the third quartile, that is, Q_3 satisfies $Pr_{x \in D}(x \leq Q_3) = 0.75$. The interquartile range $IQR = Q_3 - Q_1$ is a measure of mid-spread of the distribution. Given non-negative parameters t_1 and t_2, we can define *fences* on both ends of the distribution:

$$LF(t_1) = Q_1 - t_1 \cdot IQR,$$

$$UF(t_2) = Q_3 + t_2 \cdot IQR.$$

We prune phrases whose scores are not within the fences as the phrases with scores below the lower fence ($LF(t_1)$) are likely to be malformed and those with scores above the upper fence ($UF(t_2)$) are likely to be of common knowledge. As the distribution of scores is not symmetric around the mean, we may need to select different pruning parameters.

Algorithm 2 EnrichWithWikipedia

Input: A candidate set of key concept phrases C for a given section; Number of desired Wikipedia page links k
Output: An ordered list of top k Wikipedia page links

1: Determine set V of nodes corresponding to concept phrases in C that match a Wikipedia page title.
2: Let W be the set of all links within Wikipedia. Define $E = \{(v_1, v_2) | v_1, v_2 \in V \text{ and } (v_1, v_2) \in W\}$. Compute the graph $G = (V, E)$ thus induced by links within Wikipedia.
3: Return the top k nodes in the decreasing order of node authority scores, where we exclude any nodes corresponding to phrases enriched earlier in the textbook.

3.2 Enriching with Web Content

Given the set of key concepts for a section, our goal is to find authoritative material on the Web for the concepts and enrich the section with a few links to such authoritative material. We use Wikipedia as the source of supplementary material. In order to contain the cognitive burden on the reader, rather than adding links to all the candidate concepts, we determine up to k central concepts amongst the candidate set and augment them.

As presented in Algorithm 2, we consider only the phrases that match a Wikipedia page title, and construct the graph induced on the corresponding Wikipedia pages and links between them. We then sort the nodes in the decreasing order of node authority scores (eg., PageRank), exclude any nodes corresponding to phrases enriched earlier in the same textbook, select top k nodes, and add pointers to the corresponding Wikipedia pages. The intuition is that we are likely to find authoritative Wikipedia pages by virtue of being pointed to by many other authoritative pages. Further, given the progressive learning nature of the textbooks, it is worthwhile to exclude concepts that have already been explained earlier in the textbook.

The number of concepts, k, selected for enriching a section can be determined using the distribution of the node authority scores. The node authority scores appear to follow Zipf's ranked distribution,

$$X_r \propto r^{-1/\alpha},$$

where X_r is the value of the r^{th} ranked node authority score and α is the tail index parameter of the underlying Pareto distribution. The tail index can be estimated by regressing the log of order statistics on the log of the scores. For a desired coverage, c (say, 80%) and a limit k_0 (say, 3) on the maximum number of concepts to be shown, we obtain

$$k = \min\left(k_0, \left(c + (1-c)n^{\left(\frac{1}{\alpha}-1\right)}\right)^{\frac{\alpha}{\alpha-1}} \cdot n\right),$$

where n is the number of nodes in the induced graph. While this determination can be made empirically as well, we suggest fitting Zipf's distribution to the node authority scores as it will help to characterize the distributions over differ-

	Sciences	Social Sciences	Commerce	Mathematics
Grade IX	Science	History, Political Science		Mathematics
Grade X	Science	History		Mathematics
Grade XI		Political Science	Accountancy, Economics, Business Studies	
Grade XII	Physics	History, Sociology	Accountancy, Economics	Mathematics

Table 1: NCERT textbooks by grade and subject

ent subjects and over different grade levels, with varying α values.

4. EXPERIMENTS

We next present the results of the experiments we performed to ascertain the effectiveness of the proposed methodology.

4.1 Datasets and Tools

We apply our techniques over a corpus of high school books published by the National Council of Educational Research and Training (NCERT), India. We consider a subset of seventeen textbooks in English language from grades IX–XII, covering four broad subject areas, namely, Sciences, Social Sciences, Commerce, and Mathematics. The list of books is given in Appendix A. Table 1 provides a breakup of books by grade and subject. There are a total of 191 chapters and 1313 sections in these books. The Wikipedia articles used for augmentation were obtained from the English Wikipedia dump dated March 12, 2010.

We used off-the-shelf Stanford POS Tagger version 3.0 (with bidirectional-distsim tagger model trained on sections 0-18 of the Wall Street Journal dataset) and WordNet version 2.1 for Windows. We performed batch lookups to Microsoft Web N-gram service in July 2010.

4.2 Data Processing

Each textbook in our corpus consists of pdf files corresponding to the table of contents page and the chapters. As NCERT website contained explicit links only to the table of content pages, the corresponding chapters were obtained by parsing the table of content pages and determining links within to individual chapters. The text content of each chapter was extracted using a pdf parser and then indexed using Lucene (lucene.apache.org). There may be malformed text extracted from tables, mathematical or chemical equations, and other non-grammatical structures. We prune such malformed phrases.

The Wikipedia XML dump was parsed to extract relevant information such as page titles and internal Wikipedia links. This information was then indexed using Lucene to enable fast lookups for match on page titles and for links from a page. In our experiments, we consider only exact matches of phrases to Wikipedia page titles. If any Wikipedia page redirected to another page, the redirect link was followed till an article page was found.

4.3 Evaluation of Key Concepts

As part of evaluation of our method for determining key concepts, we first study the complementarity of WordNet and Stanford POS Tagger. We then study the effectiveness of the three grammatical patterns (P_1, P_2, and P_3) in obtaining terminological noun phrases. Finally, we investigate the effect of pruning using Microsoft Web N-gram Service.

Subject	All Nouns	Proper Nouns	Adjectives
Sciences	80.0%	82.8%	94.7%
Social Sciences	67.6%	69.0%	95.0%
Commerce	73.2%	73.3%	95.3%
Mathematics	77.5%	77.4%	94.5%
All	74.0%	74.5%	95.0%

Table 2: Coverage of WordNet over Stanford POS tags

Subject	Consistency	Disagreements uniquely corrected
Sciences	98.6%	83.9%
Social Sciences	98.5%	83.3%
Commerce	98.1%	80.2%
Mathematics	98.2%	89.4%
All	98.3%	83.4%

Table 3: Consistency between WordNet and Stanford POS Tagger

4.3.1 *Complementarity of WordNet and Stanford POS Tagger*

Our manual inspection revealed that Stanford POS Tagger makes mistakes on ill-formed sentences and sentences containing unknown words. The reason is that Stanford POS Tagger is aggressive – it returns part-of-speech tags even for unknown words and always returns a unique assignment for a sentence.

We therefore studied the coverage of Stanford tags by WordNet to explore if WordNet can be used to correct the errors made by the POS tagger. Table 2 shows that WordNet recognizes about 74% of the words tagged as nouns by Stanford POS Tagger and recognizes 95% of the adjectives. Note the variability of this statistic across different subjects; WordNet recognizes about 80% of nouns in Sciences books but only about 68% of nouns in Social Sciences books. This variability is perhaps because Social Sciences books typically contain concepts local to a region, whereas scientific terminology is universal.

Next we studied how often Stanford POS Tagger and WordNet produce the same tags. Higher consistency would imply greater confidence in the quality of our tags. As shown in Table 3, over the set of all phrases satisfying the pattern P_1, there is a very high consistency (greater than 98%). We observed similar consistency for the other two patterns.

We then analyzed the disagreements between WordNet and Stanford POS Tagger. We chose a random sample of 100 disagreements and manually checked the correctness of WordNet and Stanford POS Tagger assignments for each phrase in the sample. We found that for 85 phrases, WordNet is correct and Stanford POS Tagger is wrong; for 12 phrases, WordNet is wrong and Stanford POS Tagger is correct. Both are correct for the remaining 3 phrases. Given that WordNet is mostly correct during disagreements, we applied the correction algorithm discussed in Section 3.1.3

Subject	Overlap		
	P_1	P_2	P_3
Sciences	32.2%	24.7%	32.5%
Social Sciences	28.2%	20.8%	28.5%
Commerce	27.2%	19.4%	23.5%
Mathematics	23.4%	19.1%	27.4%
All	27.8%	21.0%	28.0%

Table 4: Comparison of linguistic patterns across different subjects

and found that it uniquely corrected nearly 83% of these phrases (Table 3).

4.3.2 Comparison of Linguistic Patterns

We compare the three grammatical patterns by studying the overlap of the generated phrases with Wikipedia page titles. Wikipedia page titles can be viewed as a large collection of concepts. We compute the overlap as the fraction of generated phrases that match a Wikipedia page title. Following [18, 19], we only use multi-word phrases in this study. We compute the overlap over each section and then average across sections from all book chapters in the corpus (or across sections from all book chapters in a subject area for subject wise analysis) and report results in Table 4.

The pattern P_1 outperforms the pattern P_2 with respect to overlap irrespective of subject area. These results reinforce the observation in [18] that P_1 outperforms P_2. The pattern P_3 performs better than the pattern P_1, albeit slightly. We found that these results continue to hold after Web N-gram based pruning. Henceforth we use P_3 for the rest of experiments.

The reader might view overlap numbers as low. Several factors play a role in determining the range of observed overlap values. For many key concepts, the corresponding Wikipedia pages might not yet have been written (eg., Tamil literary work, "Thiruvembavai"). Another factor is the choice of the matching algorithm. We perform exact matches, thereby not matching closely related or synonymous Wikipedia articles (eg., "Brihadishvara temple" vs "Brihadeeswarar temple", which is present in Wikipedia). Further a candidate concept phrase may be very specific (due to maximal pattern match) whereas there may be a Wikipedia article on a more general topic (eg., "Mature Harappan culture" vs "Harappan culture", which is present in Wikipedia). Finally, note that what matters is being able to narrow down the top candidate concepts well irrespective of whether all of them have corresponding Wikipedia pages since we might be able to augment them with other web sources.

4.3.3 Benefit of Web N-gram based Pruning

We next understand the effect of pruning using the Web N-grams, by studying the following two metrics:

Change in overlap. A way to evaluate the pruning algorithm is to measure the improvement in overlap after pruning.

Pruning rates. The pruning results in exclusion (or *drop*) of some of the candidate phrases. The dropped phrases can be of two types, those matching Wikipedia page titles (*bad drops*) and those that do not (*good drops*). The initial set of phrases can be partitioned into *covered phrases* (those that match Wikipedia page titles) and *uncovered phrases*. We define Rate of Undesirable Pruning (UPR) as the ratio of *bad drops* to *covered phrases*, that is, the fraction of covered phrases that get dropped. Similarly we define Rate of Desirable Pruning (DPR) as the ratio of *good drops* to *uncovered phrases*, that is, the fraction of uncovered phrases that get dropped. It is desirable to have low UPR and high DPR.

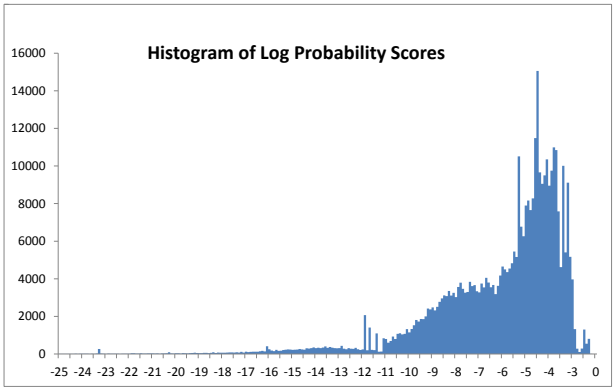

Figure 2: Histogram of log probability scores obtained from Web N-gram service

Subject	Pruning Rate		Overlap	
	Desirable (DPR)	Undesirable (UPR)	Initial	After pruning
Sciences	24.3%	0.2%	32.5%	38.7%
Social Sciences	26.0%	0.2%	28.5%	34.7%
Commerce	28.0%	0.5%	23.5%	29.3%
Mathematics	22.2%	0.1%	27.4%	32.3%
All	25.1%	0.2%	28.0%	33.8%

Table 5: Benefit of pruning for P_3 across different subjects

Pruning parameters: Figure 2 shows the histogram of log probability scores obtained from Web N-gram service, over all candidate phrases for the pattern P_3. To illustrate the histogram better, we omit a small fraction of phrases with log probability score below -25. We notice that the distribution is heavily skewed, suggesting the use of different pruning parameters for lower fence and upper fence. After some experiments where we balance between UPR and DPR, we set pruning parameters to $t_1 = 1$ and $t_2 = 0$. The quartile values are $Q_1 = -7.6$ and $Q_3 = -4.1$ and the corresponding fence values are $LF = -11.1$ and $UF = -4.1$.

Improvement with pruning: Table 5 shows the benefit of pruning for the pattern P_3 across different subjects. We observe that pruning using Web N-grams helps to improve overlap, while incurring small rate of undesirable pruning, suggesting that Web N-grams can be quite effective in identifying phrases that are malformed or represent common concepts.

4.4 Quality of Enrichments

After computing the candidate set of concepts, we use Algorithm 2 to obtain the final phrases. Once we match a phrase with a Wikipedia page, we use the page title as the concept phrase. The advantage is that we obtain the canonical form of concepts due to Wikipedia redirects. For example, "Harappan culture" redirects to "Indus Valley civiliza-

tion". We use PageRank as the node authority score in step 3. Further, while selecting the top k nodes in step 3, we restrict to nodes corresponding to multi-word phrases since we observed that Wikipedia pages with canonical single word titles typically corresponded to broad concepts. However such pages would contain pointers to more specific concepts. We therefore include these pages in the graph construction to incorporate these endorsements. For illustration, we show the graph for the Introduction section of Chapter 8 of Grade XII Physics book in Figure 3.

In order to evaluate the quality of enrichments, we selected three sections each from the four subject areas. These sections came from five books. Some sections were from initial chapters, some from the middle, and some from later chapters.

Table 6 shows the top three augmentations for each of the sections. An augmentation consists of the hyperlinked title of the Wikipedia page. The reader can verify that we are able to provide meaningful and relevant enrichments to each section.

The augmentations presented in Table 6 do not take into account progressive learning nature of textbooks. Table 7 illustrates the effect of incorporating sequentiality of sections in generating the augmentations. The results are shown for Chapter 9 of Grade XII Physics book. The left column shows the augmentations without considering sequentiality, whereas the right column shows the augmentations after accounting for sequentiality by excluding phrases enriched earlier in the chapter. The italicized augmentations on the left are redundant and are replaced by the italicized augmentations on the right. We thus see that incorporating sequentiality enhances the enrichments of the book.

5. CONCLUSIONS AND FUTURE WORK

Recognizing the crucial role of education in development and the importance of the textbooks in creating a high quality education system, we focused on devising technology for enriching textbooks with authoritative web content. We augment textbooks at the section level for key concepts discussed in the section that would benefit most from enrichment. We use ideas from data mining for identifying the concepts as well as for determining the authoritative content to be used for augmentation. We evaluated the proposed techniques using textbooks published by the National Council of Educational Research and Training (India). This evaluation shows that we are able to enrich textbooks on different subjects and across different grades with high quality augmentations using automated techniques.

During the course of this work, we obtained results that could be of general interest. These include:

1. Concepts in our system correspond to terminological noun phrases. The Natural Language Processing (NLP) community frequently employs two patterns proposed in [18] for identifying terminological noun phrases within a body of technical text. Applying these patterns to a different application domain and corpora, we confirmed that the pattern $P_1 = C^*N$ outperforms the pattern $P_2 = (C^*NP)^?(C^*N)$. We also identified another pattern, $P_3 = A^*N^+$, which exhibited slightly better performance than P_1 in our application. Here N refers to a noun, P a preposition, A an adjective, and $C = A|N$.

Physics (XII, Physics)
Chapter 8: Electromagnetic Waves
Section: Introduction
Magnetic field
Electric current
Electric field
Chapter 9: Ray Optics and Optical Instruments
Section: Refraction
Refractive index
Focal length
Lens (optics)
Chapter 15: Communication Systems
Section: Modulation and its Necessity
Alternating Current
Electromagnetic Radiation
Angular Frequency
Introductory Macroeconomics (XII, Economics)
Chapter 1: Introduction
Section: Emergence of Macroeconomics
John Maynard Keynes
Great Depression
Good (economics)
Chapter 5: The Government: Functions and Scope
Section: Fiscal Policy
Government debt
Aggregate Demand
Share (finance)
Chapter 6: Open Economy Macroeconomics
Section: Trade Deficits, Savings and Investment
Central bank
Exchange rate
Gold Standard
Themes In Indian History (XII, History)
Chapter 7: An Imperial Capital Vijayanagara
Section: Rayas, Nayaks and Sultans
South India
Telugu Language
Vijayanagara Empire
Chapter 11: Rebels and the Raj
Section: What the rebels wanted
God (disambiguation)
Major religious groups
Mughal Empire
Indian Society (XII, Sociology)
Chapter 4: The Market as a Social Institution
Section: Understanding Capitalism as a Social System
Good (economics)
Final goods
Karl Marx
Mathematics (XII, Mathematics)
Chapter 7: Integrals
Section: Introduction
Differential calculus
Fundamental theorem
Chapter 13: Probability
Section: Random Variables and its Probability Distributions
Random variable
Real number
Probability distribution
Chapter 9: Differential Equations
Section: Methods of Solving First Order,
First Degree Differential Equations
Differential equation
Cartesian coordinate system
Linear differential equation

Table 6: Augmentations across different subject categories

Physics Grade XII	
Chapter 9: Ray Optics and Optical Instruments	
Section: Introduction	
Electromagnetic radiation René Descartes Electromagnetic spectrum	Electromagnetic radiation René Descartes Electromagnetic spectrum
Section: Reflection of light by Spherical Mirrors	
Line (Geometry) Lens (optics) Ray (optics)	Line (Geometry) Lens (optics) Ray (optics)
Section: Refraction	
Refractive index Focal length *Lens (optics)*	Refractive index Focal length *Corrective lens*
Section: Total Internal Reflection	
Refractive index Optical fiber Total internal reflection	Optical fiber Total internal reflection *Laser pointer*
Section: Refraction at Spherical Surfaces	
Refractive index *Lens (optics)* *Focal length*	*Real image* *Visible spectrum* *Virtual image*

Table 7: Augmentations after incorporating the sequentiality of sections (first five sections of Chapter 9 of the book)

2. We showed that Web N-grams can be quite effective in identifying concepts that are malformed or represent common knowledge concepts.

3. We demonstrated that a lexical database such as WordNet can be quite complementary to a generic part-of-speech tagger such as Stanford POS Tagger. WordNet being a human curated system generally has better accuracy but lower coverage. We were able to successfully use WordNet for correcting errors made by the POS tagger.

The work presented here opens up opportunities for future research along several dimensions:

1. Our component for identifying key concepts employs off-the-shelf Stanford POS tagger. It will be useful to quantify the loss in accuracy from not using the application-specific corpora for training the tagger. It will also be useful to quantify the downstream effect of the tagger errors if the WordNet corrections were not applied. This information will be particularly valuable for applying techniques from this paper to textbooks written in languages other than English where tag training-sets and lexical resources like WordNet might not be so well developed. Similarly, it will be instructional to explore the use of light parsing strategies such as chunk parsing [5].

2. The books in our text corpus did not have indices. It will be interesting to investigate how the results would be affected if one were to use indices from another corpus as proxies for key concepts.

3. Another interesting direction would be designing an evaluation methodology and performing a large user study to assess the quality of enrichments. For example, two groups of users could be asked to answer a set of questions by referring to the original and the augmented texts. If the augmentation does indeed improve the text book, we should notice an improvement in the performance of the users. One can also take cues from the methodologies used in the user-studies for studying the readability of books, e.g. [13].

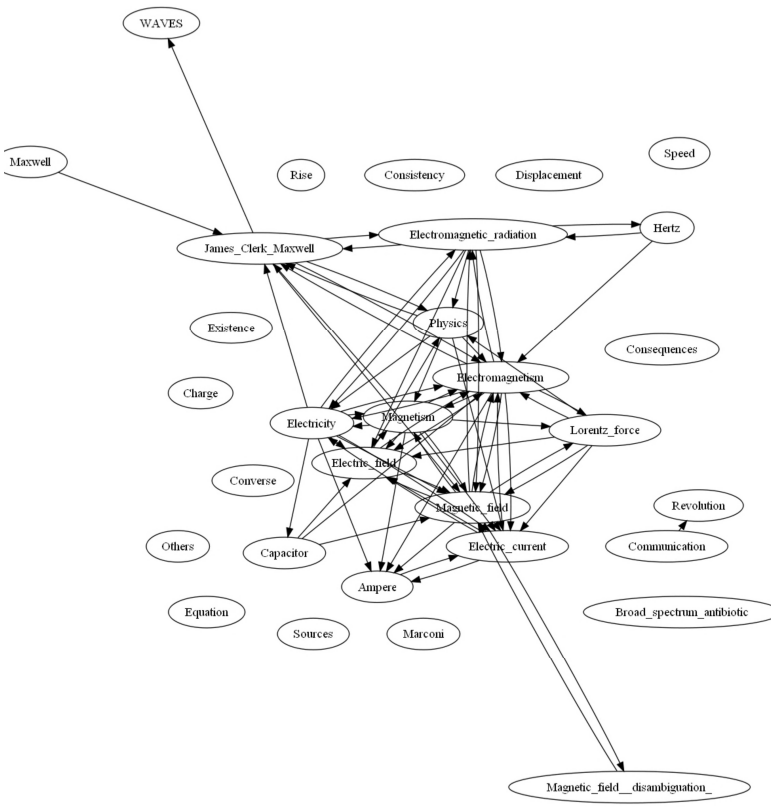

Figure 3: Wikipedia induced graph for a section in a Grade XII Physics book

4. Yet another direction would be studying issues in deploying the proposed system, taking into account factors such as school ecology and how the augmented material is made available to students and teachers.

6. ACKNOWLEDGMENTS

We are grateful to Gordon Bell, Ron Kaplan, Patrick Pantel, Partha Talukdar, Scott Waterman, and the anonymous reviewers for insightful comments and feedback.

7. REFERENCES

[1] *Knowledge for Development: World Development Report 1998/99*. World Bank, 1998.
[2] *Public Report on Basic Education in India, The PROBE team*. Oxford University Press, 1999.
[3] Improving India's education system through information technology. *IBM*, 2005.

[4] Remarks of Bill Gates, Harvard Commencement 2007. *Harvard Gazette*, 2007.
[5] S. Abney. Parsing by chunks. *Principle-based parsing*, pages 257–278, 1991.
[6] J. P. G. Chimombo. Issues in basic education in developing countries: An exploration of policy options for improved delivery. *Journal of International Cooperation in Education*, 8(1), 2005.
[7] M. Crossley and M. Murby. Textbook provision and the quality of the school curriculum in developing countries: Issues and policy options. *Comparative Education*, 30(2):99–114, 1994.
[8] C. Fellbaum. *WordNet: An electronic lexical database*. MIT Press, Cambridge, MA, 1998.
[9] B. Fuller. What school factors raise achievement in the Third World? *Review of educational research*, 57(3):255–292, 1987.
[10] B. Fuller and P. Clarke. Raising school effects while ignoring culture? Local conditions and the influence of classroom tools, rules, and pedagogy. *Review of educational research*, 64(1):119–157, 1994.
[11] J. Gillies and J. Quijada. Opportunity to Learn: A high impact strategy for improving educational outcomes in developing countries. *USAID Educational Quality Improvement Program (EQUIP2)*, 2008.
[12] P. Glewwe, M. Kremer, and S. Moulin. Many Children Left Behind? Textbooks and Test Scores in Kenya. *American Economic Journal: Applied Economics*, 1(1):112–35, January 2009.
[13] R. Guillemette. Predicting readability of data processing written materials. *ACM SIGMIS Database*, 18(4), 1987.
[14] E. A. Hanushek and L. Woessmann. The role of education quality for economic growth. *Policy Research Department Working Paper 4122, World Bank*, 2007.
[15] S. Heyneman, J. Farrell, and M. Sepulveda-Stuardo. Textbooks and achievement in developing countries: What we know. *Journal of Curriculum Studies*, 13(3), 1981.
[16] S. Heyneman, D. Jamison, and X. Montenegro. Textbooks in the Philippines: Evaluation of the Pedagogical Impact of a Nationwide Investment. *Educational Evaluation and Policy Analysis*, 6(2):139–150, 1984.
[17] D. Jurafsky and J. Martin. *Speech and language processing*. Prentice Hall, 2008.
[18] J. S. Justeson and S. M. Katz. Technical terminology: some linguistic properties and an algorithm for indentification in text. *Natural Language Engineering*, 1(1), 1995.
[19] B. Lent, R. Agrawal, and R. Srikant. Discovering trends in text databases. In *KDD*, 1997.
[20] L. Lessig. *Code and other laws of cyberspace*. Basic Books, 1999.
[21] M. Lockheed and E. Hanushek. Improving educational efficiency in developing countries: What do we know? *Compare: A Journal of Comparative and International Education*, 18(1):21–38, 1988.
[22] M. Lockheed, S. Vail, and B. Fuller. How textbooks affect achievement in developing countries: Evidence from Thailand. *Educational Evaluation and Policy Analysis*, 8(4):379–392, 1986.
[23] P. Menon. Mis-oriented textbooks. *Frontline*, August 2002.
[24] R. Mohammad. Practical constraints upon teacher development in Pakistani schools. *Professional Development in Education*, 30(1):101–114, 2004.
[25] R. Mohammad and R. Kumari. Effective Use of Textbooks: A Neglected Aspect of Education in Pakistan. *Journal of Education for International Development*, 3(1), 2007.
[26] J. Moulton. How do teachers use textbooks and other print materials: A review of the literature. *The Improving Educational Quality Project, South Africa*, 1994.
[27] J. Oakes and M. Saunders. Education's most basic tools: Access to textbooks and instructional materials in California's public schools. *Teachers College Record*, 106(10), 2004.
[28] D. Paranjpe. Learning document aboutness from implicit user feedback and document structure. In *Proceeding of the 18th ACM conference on Information and knowledge management*, 2009.
[29] D. Pennycuick. School Effectiveness in Developing Countries - A Summary of the Research Evidence. *Education Research Papers*, 1993.
[30] B. Ribeiro-Neto, M. Cristo, P. Golgher, and E. Silva de Moura. Impedance coupling in content-targeted advertising. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 2005.
[31] A. Riddell. *Factors influencing educational quality and effectiveness in developing countries: A review of research*. Deutsche Gesellschaft fur Technische Zusammenarbeit (GTZ), Germany, 2008.
[32] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL-HLT*. Association for Computational Linguistics, 2003.
[33] K. Wang, C. Thrasher, E. Viegas, X. Li, and P. Hsu. An overview of Microsoft Web N-gram corpus and applications. In *NAACL-HLT*. Association for Computational Linguistics, 2010.
[34] W. Yih, J. Goodman, and V. Carvalho. Finding advertising keywords on web pages. In *WWW*, 2006.

APPENDIX

A. LIST OF NCERT TEXTBOOKS IN OUR CORPUS

1. Sciences
   (a) Science - Textbook for Class 9
   (b) Science - Textbook for Class 10
   (c) Physics - Textbook for Class 12
2. Social Sciences
   (a) India and the Contemporary World-I, A textbook in History (Class 9)
   (b) Democratic Politics (Class 9)
   (c) India and the Contemporary World-II, A textbook in History (Class 10)
   (d) Political Theory (Class 11)
   (e) Themes in Indian History (Class 12)
   (f) Indian Society (Class 12)
3. Commerce
   (a) Accountancy - Textbook for Class 11
   (b) Business Studies - Textbook for Class 11
   (c) Indian Economic Development (Class 11)
   (d) Accountancy - Textbook for Class 11
   (e) Introductory Macroeconomics (Class 12)
4. Mathematics
   (a) Mathematics for Class 9
   (b) Mathematics for Class 10
   (c) Mathematics for Class 12

[Note that class is the term used to refer to grade in India.]