Highlight Research Outline

David Y.J Kim

December 1, 2020

This is an outline how I am planning to further my research after "Inferring Student Comprehension from Highlighting Patterns in Digital Textbooks: An Exploration in an Authentic Learning Platform"[1]. None of this work has been published and it's writing is rough since the main purpose of this outline is to aid my research

1 Introduction and Data

The overall research project focus on developing intelligent textbooks which infer student understanding based on students' interactions with the textbooks. Given a model of the student's knowledge state, we then hope to customize material for further study and review.

Data were collected in collaboration with OpenStax, a nonprofit organization that supports open-access college-level digital textbooks. For two full semesters data were collected in Biology, Physics, Sociology, and History classes with 11,134 students. In these classes, students were able to highlight and add annotations to their e-textbooks while reading. Of the 11,134 students, 2,829 used the highlighting facility. Given the record of these highlights and annotation, we attempt to infer student comprehension, as assessed by a quiz that students take at the end of each section as well as delayed review questions administered about a week after initial reading.

Metadata of the data: Each textbook is divided into chapters which are further subdivided into *sections*

- Textbook : Total of 6 textbooks was used for this data. Some AP course textbooks were added in the data which are for high school students.
  - Biology : Biology for AP, Biology 2e
  - Physics : Physics for AP, College Physics with Courseware
  - Sociology : Introduction to Sociology
  - History : BRI APUSH(AP US History)
- Sections: The basic unit of the paragraph for this analysis. There were a total of 1011 distinct sections
  - Biology
    - The number of sections in Biology 2e is 255
    - The number of sections in Biology for AP Courses is 125
  - Physics
    - The number of sections in College Physics with Courseware is 284
    - The number of sections in College Physics for AP 2e is 144
  - Sociology
    - The number of sections in Introduction to Sociology 2e is 169
  - History
    - The number of sections in BRIAPUSH is 33
- Sessions : There were a total 830354 number of sessions(student + section), from these sessions 27019 of them had highlights

- Biology
    - Biology 2e : total of 342693 sessions and 11396 of them have highlights
    - Biology for Ap Courses : total of 21040 sessions and 5559 of them have highlights
- Physics
    - Physics with Courseware : total of 384672 sessions and 7695 of them have highlights
    - Physics for AP : total of 39143 sessions and 990 of them has highlights
- Sociology
    - Introduction to Sociology : total of 42683 sessions and 1377 of them has highlights
- History
    - BRI APUSH : total of 89 sessions and 2 of them has highlights

The goal is to predict quiz performance from the pattern of highlighting. Previously[1], we built separate linear models for each section of text. The input data used for prediction is a high dimensional binary feature vector where each feature indicates whether or not a given word of a section is highlighted. Finding the best representation of highlights as input to the linear model was the first task. We found that parsing the whole passage into words and reducing the dimension using PCA (Principal Component Analysis) can explain about 13% variance of the test performance. Although, this is an exciting result, this method is extremely limited in that it is text dependent: we build a separate model for each section of text and therefore require training data for each section. We need to further explore whether we can build a section-independent model that uses the content of the text and the content of the quiz questions to predict student performance.

2 Methodology

Our initial attempt is try using BERT (Bidirectional Encoder Representations from Transformers)[2], a well known Natural Language Process/Deep Learning technique. Further research[3] has found that it is possible to find sentence pairs with the most similar semantic meanings. Using this idea, we can feed the highlighted sentence along with the questions and see if students who highlights phrases related to the question have a higher chance of getting the question correct.

There are still some obstacles that needs to be addressed. First of all, we have to decide how to partition the textbook into segments(words, phrases, sentences etc.). From here and on we will use the notification $B(s, q)$ for the BERT match score between segment s and question q, and N for the number of segments. Second, we need to determine a representation of how much of segment s is highlighted by the student. We will call this representation H_s. We can consider alternatives like:

$$H_s^1 = \begin{cases} 1 & \text{if any word of the segment is highlighted} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$H_s^2 = \frac{\text{words in the segment that are highlighted}}{\text{number of words in the segment}} \tag{2}$$

where $s \in S$, S is the full set of segments in a section, $|S| = N$

Now we have to decide how we will deal with the multiple segments students have highlight for the section. For example, if a student highlights several sentences should we only consider the max related highlight with respect to the question? Or should we consider the average relation?

We are planning to approach this in a brute force manner to see which methods comes up with the best result. Our approach is from starting from maximum ($max(B(s, q)),\ \ s \in S$) toward the average ($\frac{1}{N} \sum_s B(s, q)$).

Notice that $\frac{1}{n}||x||_1$ is the same as the arithmetic mean when $x \in R^n$. Also, $||x||_\infty$ is the same as the maximum element in x. Also, norms have a relation such as the following $||x||_r \leq n^{\frac{1}{r}-\frac{1}{p}}||x||_p$. If we apply this rule as $p = r + 1$ for all $r = 1, 2...$, then we get the following relation.

$$||x||_1 \leq n^{\frac{1}{2}}||x||_2 \leq n^{\frac{2}{3}}||x||_3 \leq ... \leq n||x||_\infty$$

Since we know that $n > 0$ we can also infer the following relation

$$\frac{1}{n}||x||_1 \leq \frac{n^{\frac{1}{2}}}{n}||x||_2 \leq \frac{n^{\frac{2}{3}}}{n}||x||_3 \leq ... \leq ||x||_\infty$$

As we can see, starting from the arithmetic average to the maximum everything between them indicates something in the middle. We are planning to choose some set of random integers and see which range of integers show the best results.

Starting from $p = 1$ to $p = \infty$ and using the definition of norm, we will obtain the following value

$$Match_{p,q} = \left[\frac{1}{N}\sum_s B(s,q)^p\right]^{1/p}$$

Now we have to determine how $B(s, q)$ and H_s are combined to predict the performance on question q. One way is replacing N for H_s as the following

$$\text{HighlightedMatch}_{p,q} = \left[\frac{\sum_s H_s B(s,q)^p}{\sum_s H_s}\right]^{\frac{1}{p}}$$

For the binary code(1), this will just compute the L_p mean match over the highlighted sections, as for the continuous code(2), this will compute a weighting that is based on H_s. (need further verification)

Von Restorff effect suggests that highlighting might allow the possibility of adversely affecting the retention of the non-highlighted material[4]. Thus, we intend to consider the phrases that were not highlighted. We could compute a second match score for the non-highlighted sections:

$$\text{NonHighlightedMatch}_{p,q} = \left[\frac{\sum_s (1-H_s) B(s,q)^p}{\sum_s 1 - H_s}\right]^{\frac{1}{p}}$$

Now we can determine an overall match:

$$\text{OverallMatch}_{p,q} = \text{HighlightedMatch}_{p,q} - \alpha \text{ NonHighlightedMatch}_{p,q}$$

From the above equation We need to determine how to get the best performance from the choice of:

- p
- α
- how to represent the highlights H_s (binary or continuous or anything else)
- how to segment the text (words, phrases, sentences)

The determination of the best performance will be based on logistic regression models to estimate performance. We compare two distinct[5], yet related models:

$$
\begin{aligned}
M_1 &: \Phi(P(Y_{q,s}) = \theta_s - \mu_i \\
M_2 &: \Phi(P(Y_{q,s}) = \theta_s - \mu_i + \eta \, \text{OverallMatch}_{p,q}
\end{aligned}
$$

Model M_1 corresponds to our base-effect model in that it does not include any effects from student highlighting. In this Model, $\Phi(p) = log\frac{p}{1-p}$ denotes the logistic function, $Y_{s,q}$ denotes the correctness of the response of student s on question q, θ_s denotes the latent ability of student s, and μ_q denotes the difficulty of question q. We note that M_1 correspond to a 1PL Item Response Theory model. Model M_2 is identical to M_1 with the exception of the last term, $\eta OverallMatch_{p,q}$. η denotes the impact that a highlight will have on the final performance outcome.

After fitting the data to both models, we then determine which model provide the better fit. To balance model complexity and fidelity, as well as to avoid overfitting, we compute the deviance information criteria (DIC) for each model. One limitation of this analysis is that it is difficult to get a sense of how large the effect of highlighting was. Thus, we will also estimate the effect size of highlighting by converting the η values for each subject to a normal effect deviate (NED) similar to Cohen's-d. As an alternative to effect sizes, we plan to examine the magnitude of the highlighting effect in terms of it's prediction. We will obtain this by computing the estimated success probabilities for each student/question pair for the case when a student made a highlight and when they did not and averaging over all observations.

$$
\begin{aligned}
\Delta &= \frac{1}{|\{S, I\}|} \sum_{s,i} P(Y_{s,i} | h_{s,i}) - P(Y_{s,i} | h^c_{s,i}) \\
&= \frac{1}{|\{S, I\}|} \sum_{s,i} \Phi(\theta_s - \mu_i + \eta) - \Phi(\theta_s - \mu_i)
\end{aligned}
$$

References

[1] David Young-Jae Kim and A. Winchell and A. Waters and Phillip J. Grimaldi and Richard Baraniuk and M. Mozer *Inferring Student Comprehension from Highlighting Patterns in Digital Textbooks: An Exploration in an Authentic Learning Platform.* iTextbooks@AIED, 2020.

[2] Jacob Devlin and Ming-Wei Chang and Kenton Lee and Kristina Toutanova *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* 2018

[3] Nils Reimers and Iryna Gurevych *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*, 2019.

[4] Fowler, R. L., Barker, A. S. (1974). *Effectiveness of highlighting for retention of text material* Journal of Applied Psychology, 59(3), 358–364. https://doi.org/10.1037/h0036750

[5] Waters, Andrew E and Grimaldi, Phillip J and Baraniuk, Richard G and Mozer, Michael C and Pashler, Harold *Highlighting Associated with Improved Recall Performance In Digital Learning Environment*