Computationally Identifying Confusing Passages in Educational Materials

Amy X. Zhang
MIT CSAIL
axz@mit.edu

Abstract

A common way to learn about new subjects is to read textbooks. manuals, or other expository materials. However, it can be difficult for experts writing the materials to write in a way that will teach novices new concepts without causing confusion. In this work, we explore different linguistic aspects of passages in educational materials that may lead to confusion. To this end, we build a corpus of textbook paragraphs taken from a physics textbook that has been annotated by several classes of physics students with their comments. Focusing on the annotations that express confusion, we seek to distinguish the passages that lead to confusion among students. We find that features around when and how new terminology is introduced, the affective and cognitive language used, and the other items surrounding the text on the page, are predictive of confusing text, among others. Our best model using linguistic features outperforms unigram and random baselines by around 20%, achieving an average of around 73% accuracy on our textbook dataset.

1 Introduction

One of the key issues when writing educational materials is to write clearly and unambiguously to students so that they do not get confused. However, writing learning materials such as textbooks that teach complex concepts can be a difficult process. The author is usually a seasoned expert in the topic and understands the material deeply, but they must also explain often complicated concepts to people encountering the information for the first time. This can be challenging when for instance, the authors do not realize they are taking prior knowledge for granted that learners do not understand or do not explicitly state certain things that learners then infer incorrectly.

Traditional educational literature has dedicated research and dispensed practical guidance around how to write good expository text (Alley, 1996). Today, with a proliferation of MOOCs and also traditional courses with online components, we have a way to investigate this question empirically. Thanks to new educational software that allows students to annotate educational materials while they are reading them, we can find out what specific parts of the text are confusing to students.

In the following sections, we describe background work on textbook analysis and related work in computational analysis of text quality. After explaining in more detail how we selected and cleaned our dataset, we examine the characteristics that differentiate confusing passages from not confusing ones. We then go on to outline a prediction task to predict the confusingness of a passage and the performance of our best model comprised of linguistic and structural features, achieving around a 20% improvement in accuracy and F1 over bag-of-words and random models.

2 Related Work

2.1 Textbooks and Good Pedagogical Writing

Much of the traditional literature around how to write effective pedagogical materials offer practical advice towards better scientific writing (Alley, 1996) or writing for textbooks. For instance, some

advise writers to add instructional objectives, statements of actions that students should be able to perform if they mastered the material (Gronlund, 1995), with actions taken from Bloom's taxonomy of educational objectives (Bloom et al., 1956).

Other work look more closely at the content of textbooks to examine their efficacy. Some worked demonstrated that concepts that take on multiple meanings, such as the concept "gene" in biology, can lead to misconceptions (Flodin, 2009). Studies have shown that inserting questions into scientific texts facilitates comprehension (Rouet and Vidal-Abarca, 2002). The inclusion of figures into text also promote systematic thinking and produces useful mental models (Mayer, 1989). Finally, studies have reported that the presence of conjunctions, overlapping terms, less ambiguous pronouns, and other explicit coherence relations in expository text leads to greater comprehension (Ozuru et al., 2009).

More recently, researchers are examining how e-textbooks are used in the case of MOOCs or courses with an online component (Nicholas et al., 2010). Researchers are also building new interactive learning materials that include dynamic presentations or automated assessment exercises (Fouh et al., 2014) or provide collaborative experiences such as discussions within the margins (Zyto et al., 2012). Importantly, the rise of online learning materials and interactivity are now allowing researchers to examine questions of effective pedagogy in a data-driven way.

2.2 Computational Analyses of Expository Text

It is possible that the question of finding confusing parts of educational text has some overlap with the literature around text readability assessment, which aims to identify the quality (Pitler and Nenkova, 2008), grade level, or difficulty of text (Feng et al., 2009). Features that have been used to effect for this problem include POS-features and average sentence length, as well as the traditional Flesch-Kincaid metrics (Feng et al., 2010). Other work looks at features that related to word concreteness, syntatic simplicity, and text cohesion, among others (Graesser et al., 2011). The research around automatic essay scoring also has overlap with regards to text readability assessment (Shermis and Burstein, 2003). While some of our features are informed by this work, our research differs from this work because we focus on educational passages that elicit confusion, which can be unrelated to the reading level of the text.

There is also some related work on detecting confusion in educational contexts such as in online MOOC forums (Yang et al., 2015) (Agrawal et al.,). However, these studies focused on detecting whether students were expressing confusion, while we explore what expository text will elicit such confusion within students.

3 Data

The data for our study comes from a software tool called Nota Bene (NB), a PDF annotation tool (Zyto et al., 2012). Teachers can choose to upload educational PDFs to NB and have students highlight parts and leave comments for the class to then discuss, effectively having discussions "in the margins". In this case, we focus on 10 chapters taken from an introductory Physics textbook, which have been annotated by many students of an introductory Physics course at a university. The course has used NB for annotation over several years, leaving different sets of annotations by different students over the same material. Overall, there are 279 pages across the 10 chapters and 18,022 initial comments left by 16 different classes or sections of a class, with 492 participants overall.

3.1 Finding Confusing Comments

Because we are interested in finding areas in the textbook chapters where students are confused, we first need to pick out only the comments where students express confusion. Since we do not have labeled data, we instead use simple heuristics to pick out the comments with confusion, such as the presence of a question mark or common phrases indicating confusion such as "unclear" or "having trouble". This lead us with 8,384 comments that were marked as expressing confusion. Manually examining a random selection of 100 of them, all 100 of the comments were found to express confusion by the author.

3.2 Finding Confusing Textbook Passages

Next, we need to find passages in the textbook that are confusing to many classrooms. From our data set of 10 chapters, 8 chapters have been annotated by 4

different classes while 2 have been annotated by 3 different classes. Looking at the annotations at the sentence level, we consider sentences that have at least one confused comment to be confusing to that class, and sentences with no confused comments to not be confusing to that class. We found that in the chapters with annotations by 3 classes, all three sets of annotations agreed on confusion of a sentence 60.6% of the time. In the chapters with annotations by 4 classes, 3 out of the 4 annotations agreed 82.9% of the time. Cronbach's alpha was 0.67 for the chapters with annotations by 3 classes and 0.71 for the chapters with annotations by 4 classes, showing acceptable reliability. Given this, we consider a sentence to be confusing if at least 3 classrooms have a confused comment highlighting that sentence, and not confusing otherwise.

Since we are interested in passages that are confusing, we then look at paragraphs in our data set. We choose paragraphs because each paragraph roughly corresponds to a high level point that the author is making, and this is the level at which we wish to gauge confusingness. We consider a paragraph to be confusing if at least 30% of the sentences in the paragraph are confusing according to the above standard. We also chose to ignore paragraphs that were part of separate boxes for "Exercise", "Checkpoint", "Self-Quiz", etc. since these were almost never highlighted by students, potentially because the course did not emphasize them. In our data set, there are 931 paragraphs in total, 595 of which we deemed confusing.

3.3 Textbook Metadata

In addition to the text of the chapters themselves, we also collect other metadata. This includes the figures and tables within a page, the page that each paragraph is on, the position of title headings and subheadings, as well as the placement and text of special boxes of content containing example exercise problems, questions for the readers, and other types of extra information. We also note when text is bolded in the textbook, for instance when a vocabulary term is first introduced as well as short summaries or key points. Finally, we extract the vocabulary words for each chapter, which is listed in the glossary.

4 Features

We next turn to considering characteristics of the paragraphs of text that correlate with confusion before turning to developing a model to predict whether a paragraph is confusing or not using these features. We explain the features as well as report the results of a Pearson correlation test with confusion for each feature.

4.1 Type of Paragraph

We first look at features which tell us a little bit about what type of paragraph the text is. For instance, if the paragraph has bolded sentences in it, then it is likely a paragraph that is attempting to summarize information into one or two succinct points. If it has one or more bolded definition term in the paragraph, then it is likely a paragraph that is laying out new points or claims to the reader. Finally, if the paragraph looks to be itemized then it is likely part of a list of actions that the author is providing to explain how to complete a task.

Feature	p	p
bolded sentences	0.230	$<$0.001
definitions	0.227	$<$0.001
part of a list	-0.116	$<$0.001

Table 1: Page Information feature correlation with confusion.

4.2 Page Information

Next, we consider features that give information about the page that the paragraph sits on. First, we have a feature for what page in the chapter the paragraph is on. We also consider the number of figures on the page, the number of tables on the page, and the number of headings on the page. Also, we collect the number of "Checkpoint" boxes on the page, which are boxes asking the reader introspective questions, and the number of "Procedure" boxes, which are boxes teaching the reader how complete a task, such as run an experiment.

4.3 Technical Information

We consider features that quantify how much technical information is in the paragraph. This includes the number of standalone equations, or equations that take up an entire line of their own with extra padding, as well as the number of equations in total,

Feature	p	p
page number	-0.374	<0.001
number of figures	0.188	<0.001
number of tables	-0.062	<0.1
number of headings	0.106	<0.005
Checkpoint boxes	0.152	<0.001
Procedure boxes	-0.138	<0.001

Table 2: Page Information feature correlation with confusion.

or equations that appear in their own line as well as inline. We also count the number of variables used in the paragraph and the number of values mentioned in the paragraph, which are well-known quantities such as Avogadro's number.

Feature	p	p
standalone equations	-0.146	<0.001
all equations	-0.210	<0.001
number of variables	-0.185	<0.001
number of values	0.024	0.472

Table 3: Technical feature correlation with confusion.

4.4 Emotion and Cognition

We also consider the type of words used in the paragraph that express affective or cognitive processes. Because we are interested in terms that generalize beyond a particular chapter's topic, we consider the relative frequency of several LIWC categories (Tausczik and Pennebaker, 2010). First we consider terms that express affect as one group, which include both positive and negative words, as well as words that express emotions such as anxiety, anger, and sadness. We also consider different words that indicate cognitive processes as a whole and also broken down into some of its components, including insight, causation, discrepancy, tentativeness, certainty. Finally, we count the occurrence of the phrases "suggest", "lead(s) us to conclude", "we can conclude", or "mention" under the category of suggestive phrases.

4.5 Vocabulary

Because we know the vocabulary terms in each chapter, we can consider how the vocabulary terms are introduced and then subsequently used to the reader. We first consider the number of vocabulary words in the paragraph. Next, we consider whether the paragraph is the first occurrence of a vocabulary word. We also count the average number of times each vocabulary word in the paragraph has been used so far in the chapter. Finally, we count the maximum number of pages since the last occurrence of one of the vocabulary words used in the paragraph.

Feature	p	p
affect	-0.108	<0.001
cogntive processes	0.021	0.525
insight	-0.096	<0.005
causation	0.044	0.183
discrepancy	-0.015	0.658
tentativeness	0.050	0.125
certainty	-0.087	<0.01
suggestive	0.116	<0.001

Table 4: LIWC feature correlation with confusion.

Feature	p	p
vocab words per sentence	0.031	0.348
first occurrence of vocab word	0.174	<0.001
times vocab words seen so far	-0.033	0.309
num pages since last vocab appearance	0.066	<0.05

Table 5: Vocab feature correlation with confusion.

4.6 Text Length

We also consider some common features from text readability regarding text length. We consider the number of sentences in the paragraph, the average number of words per sentence, the average sentence length, and the average number of characters per word.

Feature	p	p
number of sentences	-0.055	<0.1
avg words per sentence	0.124	<0.001
avg sentence length	0.177	<0.001
avg characters per word	0.188	<0.001

Table 6: Text length feature correlation with confusion.

4.7 Part-of-speech Information

Some other common text readability features include the prevalence of different parts of speech. We consider the prevalence of adjectives, numerals, and nouns from a tagged corpus generated by the Stanford POS Tagger (Toutanova et al., 2003).

Feature	p	p
adjectives	0.173	<0.001
numerals	-0.067	<0.05
nouns	-0.055	<0.1

Table 7: POS feature correlation with confusion.

4.8 Non-vocabulary Text

Finally, we consider terminology that may not been signposted and defined in the book as vocabulary words but may still be physics terms or more common terms. For the purpose of this task, we manually build a corpus of 63 terms that are common words but have specific meaning in a physics context. This includes words such as "force", "field", "ray", and "neutral" as examples. We also make use of the 10,000 most common English words taken from Google's Trillion Word Corpus [1]. First we look at all words that are nouns or adjectives in the paragraph and count their average occurrence over the 10 chapter data set to get a measure of how used the main words in the paragraph are in the chapter. We next consider whether the paragraph is using any nouns and adjectives for the very first time in the chapter. We have another feature that computes the same number but only for terms that are not in the Google common words corpus. From inspection, many of these words are more complex physics terms that were nonetheless not vocabulary in the chapter. Finally, we count the average number of common physics terms in the paragraph.

Feature	p	p
avg word occurrence	0.214	<0.001
word first occurrence	0.171	<0.001
uncommon word first occurrence	0.099	<0.005
number of common physics terms	0.216	<0.001

Table 8: Non-vocab feature correlation with confusion.

5 Results

We now develop models to predict whether a paragraph will be confusing or not. Given our data set of 931 paragraphs, we split it into training and test sets using 10 fold cross-validation and average the results. As a comparison to the linguistically informed model we build based on our features, we compute several baselines. We try several models using unigram vectors with and without tf-idf normalization. All of these models have worse accuracy than the model where we simply predict all items are not confusing.

Instead the models that we create using linguistic features that we defined earlier perform better than all the baselines across all measures. The best model is a logistic regression model that uses maximum entropy and achieves on average 73% accuracy and has an F1 score of 0.56. However, not all the features mentioned earlier made the best model improve, even though some of them were highly correlated linearly.

6 Discussion

As we saw, unigram models performed poorly in our task to predict confusion. Though they are not reported here, earlier attempts to use word vectors, including ones trained on a Google corpus and ones trained on our training data set also performed poorly. Trials using bigrams and different frequency cutoffs did not improve the models. In addition, from earlier studies using a separate Biology data set adding unigram or bigram vectors to the best linguistic model also made it perform slightly worse. One potential reason why bag-of-words models do not perform well on this task could be that because we are looking at paragraphs from a number of different chapters covering different material, the words themselves take into account too much of the actual content of the textbook. For instance, a new term introduced in one chapter may be confusing, but then it may become used frequently in subsequent chapters and no longer signal confusion because the readers have learned the concept behind the word.

When looking at our linguistic models that perform well, for instance, the MaxEnt model or the RF model, we can look at the coefficients of the model or the feature importances that the model reports. However, these numbers, along with the linear correlations described earlier, can be difficult to interpret since they don't always correspond with each other. Part of the problem may stem from the issue that many features may be correlated with each. This may lead to unexpected or unintuitive correlations with confusion. For instance, we find that

[1] https://github.com/first20hours/google-10000-english

Model	Acc	Precision	Recall	F1
Random (statified)	0.54	0.36	0.36	0.36
All Confusing	0.64	0	0	0
Unigram (SVM)	0.49	0.30	0.29	0.29
Unigram tf-idf (SVM)	0.47	0.26	0.22	0.22
Unigram (MaxEnt)	0.50	0.31	0.27	0.27
Unigram tf-idf (MaxEnt)	0.53	0.14	0.07	0.08
Unigram (Naive Bayes)	0.49	0.31	0.33	0.28
Unigram tf-idf (Naive Bayes)	0.61	0.05	0.03	0.04
Linguistic Model (SVM)	0.70	0.64	0.42	0.49
Linguistic Model (RF)	0.70	0.65	0.44	0.51
Linguistic Model (MaxEnt)	**0.73**	**0.72**	**0.50**	**0.56**

Table 9: Performance across different models.

Feature	MaxEnt	RF
bolded sentences	1.108	0.030
definitions	1.158	0.041
part of a list	-0.698	0.004
page number	-0.099	0.150
num figures	NA	0.038
standalone equations	0.150	0.013
number of variables	-0.220	0.039
discrepancy	-0.055	0.025
cognitive processes	0.006	0.065
tentative	0.023	0.036
suggestive	0.967	0.007
first occur. of vocab word	0.216	0.018
num pages since last vocab	0.002	0.023
number of sentences	-0.003	0.045
avg sentence length	0.003	0.115
adjectives	1.254	0.103
numerals	-0.034	0.084
word first occur.	NA	0.103
uncommon word first occur.	NA	0.061

Table 10: Coefficient in best MaxEnt model and feature importance in best RF model.

the number of figures on the page positively correlates with confusion, which is counter to what we might expect (Mayer, 1989). We also note that there is a moderate negative correlation between the page number of the paragraph and confusion. This might happen for a number of reasons such as students are losing interest in annotation or are too busy with other things over the course of the class. In any event, this particular feature obviously does not provide useful or generalizable information to future writers or other data sets, and may be a confounding factor to the true correlation of other features.

Other features of this nature include the features that distinguish summaries or definitions of new vocabulary. Though both these features were decent predictors of confusion, they do not provide great linguistic insight since one cannot simply introduce less vocabulary, for instance. Instead, it may be important to normalize by the type of paragraph, since certain types of paragraphs are more confusing than others. For example, paragraphs introducing new vocabulary could be more confusing on the whole, but within this group, there may be some examples that are more confusing and others that are less confusing. Distinguishing between these examples may provide deeper insight.

7 Future Work

Given the difficulties around interpretation, future work will revolve around normalizing the data so that comparisons are between like paragraphs. For instance, we can change the prediction task so that it is about predicting from pairs of confusing and not-confusing passages that are on the same page. Additionally, we can restrict the pairs so that only paragraphs serving the same function (explanations, definitions, summaries, etc.) are compared. Classifying paragraphs into their respective functions with regards to the purpose of the learning materials is in itself a difficult task, without labeled data. In this case, bag-of-words models might actually perform well, since the types of words used between explanations and definitions can be very different. Some initial attempts at labeling paragraphs have revealed that sometimes paragraphs have a mix of functions, which means we may also need to segment the paragraphs into smaller chunks as well. Unsupervised

Bayesian topic segmentation might be useful for this purpose (Eisenstein and Barzilay, 2008).

Since our classification strategy left us with many paragraphs (almost 40%) that were deemed confusing, non-binary methods such as regression or multi-class classification that can get at the degree or ranking of confusion level would be useful for writers to prioritize their attention.

Another interesting line of future work is around how to incorporate the annotation discussion data further. For instance, many of the posts expressing confusion have follow-up comments that answer the students' questions or provide further useful information. This sort of information could be helpful for providing suggestions for authors that they could incorporate into their text. For classrooms that already use NB, the features that we use and additional features taken from the post data and incorporated into our model could help direct the teachers of the class to areas where they should intervene.

Additionally, it would be interesting to segment students providing the annotations into different categories. For example, one could split the students into the ones that achieved high grades and the ones that achieved low grades. Differences in what the well-performing versus poor-performing groups find confusing would be very interesting to study.

8 Conclusion

Thanks to new annotation technology, we can now find out where exactly in educational texts students are confused. This presents all sorts of interesting possibilities around understanding the nature of confusion and building tools to help teachers and writers. In this work, we explore the linguistic and structural aspects of an educational text that has been annotated by students to be able to predict which paragraphs are likely to be confusing. Our features include aspects of the text such as the introduction of new vocabulary and other terms, the affective and cognitive words used in the text, as well as the other items on the page. Our results outperform several baseline models by around 20%, achieving accuracy around 70%.

9 Code

The code for this work can be found at the public github repository: https://github.mit.edu/axz/nb_project.

References

Akshay Agrawal, Jagadish Venkatraman, Shane Leonard, and Andreas Paepcke. Youedu: Addressing confusion in mooc discussion forums by recommending instructional video clips. In *Educational Data Mining*.

Michael Alley. 1996. *The craft of scientific writing*. Springer Science & Business Media.

Benjamin Samuel Bloom, Committee of College, and University Examiners. 1956. *Taxonomy of educational objectives*, volume 1. David McKay New York.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 334–343. Association for Computational Linguistics.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 229–237. Association for Computational Linguistics.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 276–284. Association for Computational Linguistics.

Veronica S Flodin. 2009. The necessity of making visible concepts with multiple meanings in science education: The use of the gene concept in a biology textbook. *Science & Education*, 18(1):73–94.

Eric Fouh, Ville Karavirta, Daniel A Breakiron, Sally Hamouda, Simin Hall, Thomas L Naps, and Clifford A Shaffer. 2014. Design and architecture of an interactive etextbook–the opendsa system. *Science of Computer Programming*, 88:22–40.

Arthur C Graesser, Danielle S McNamara, and Jonna M Kulikowich. 2011. Coh-metrix providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5):223–234.

Norman Edward Gronlund. 1995. *How to write and use instructional objectives*. Simon & Schuster Books For Young Readers.

Richard E Mayer. 1989. Systematic thinking fostered by illustrations in scientific text. *Journal of educational psychology*, 81(2):240.

David Nicholas, Ian Rowlands, and Hamid R Jamali. 2010. E-textbook use, information seeking behaviour and its impact: Case study business and management. *Journal of Information Science*, 36(2):263–280.

Yasuhiro Ozuru, Kyle Dempsey, and Danielle S McNamara. 2009. Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and instruction*, 19(3):228–242.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 186–195. Association for Computational Linguistics.

Jean-François Rouet and Eduardo Vidal-Abarca. 2002. Mining for meaning: Cognitive effects of inserted questions in learning from scientific text. *The psychology of science text comprehension*, pages 417–436.

Mark D Shermis and Jill C Burstein. 2003. *Automated essay scoring: A cross-disciplinary perspective*. Routledge.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 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*, pages 173–180. Association for Computational Linguistics.

Diyi Yang, Miaomiao Wen, Iris Howley, Robert Kraut, and Carolyn Rose. 2015. Exploring the effect of confusion in discussion forums of massive open online courses. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, pages 121–130. ACM.

Sacha Zyto, David Karger, Mark Ackerman, and Sanjoy Mahajan. 2012. Successful classroom deployment of a social document annotation system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1883–1892. ACM.