Textbook to Bullet Points

Shreya Nakkala
Texas A&M University
College Station, TX USA

Melissa Zhang
Texas A&M University
College Station, TX USA

Abstract

Textbooks can be dense with words pertaining to a cluster of topics. Reading and searching through such a wall of text is difficult and tedious. In order to create a better reading experience for the reader, we propose splitting the text up into individual sentences which are bullet pointed, grouped based on topic, and indented based on whether the sentence is introducing the topic, is a detail of the topic, or is an example for the topic.

Topic groupings will be achieved by using Sequential Latent Dirichlet Allocation to identify the topics in individual paragraphs. A Convolutional Neural Network would then be applied to the grouped sentences to classify whether the sentence is a main sentence, detail, or example. The final classification will then determine the indentation applied to the sentence.

1. Introduction

Textbooks are an essential part of a student's educational life. Yet, textbooks are not always the easiest to read. Some common reasons for this difficulty in reading was cited to be due to an overload of information and the inability to recognize text structures such as the relationship between main and sub topics [6]. Our proposal then is to break the text up into individual sentences which are bullet pointed, grouped by topic, and indented based on what purpose the sentence served for the topic. By doing so, our goal is to break up the text within the textbook into more readable segments which can show a very basic relationship between the sentences. We also propose to not remove any sentences in order to prevent a loss of information. The final result would then be a reconstruction of the textbook formatted as a rough outline.

Our method would first parse through a paragraph and clean the text of stop words. Then, using Sequential Latent Dirichlet Allocation which takes in the sequence of sentences, we can gather what topics exist in the given paragraph. Given the topics, we can group each sentence based on the main topic for the sentence. Each sentence in the sentence groups will then classify whether each of the sentences is a main idea, detail, or example. The main idea will be classified as the first sentence for a topic. Then the remaining sentences will go through a Convolutional Neural Networks which will classify the sentence as a detail or example. The final classification will then determine the indentation of sentences, providing an organized and bullet pointed outline for the specific section.

2. Preliminary Literature Survey

Due to the need of analyzing natural text within a textbook, our research looked towards natural language processing which is the theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts [4]. Specifically, we focused on topic modeling which can mine topics from text corpora [5].

One topic modeling technique we considered using was Latent Dirichlet Allocation (LDA) which has been ruled the topic modeling technique since its inception [5]. Latent Dirichlet Allocation is a probabilistic topic model which contains the basic idea that documents represent a random mixture of latent topics. Each topic is then characterized by a distribution over words [1]. However, due to LDA's bag of words approach at topic modeling, we lose important information about the sequence of topics.

Another topic modeling technique we examined was the Sequential Latent Dirichlet Allocation (seqLDA) which is a variation of LDA. Unlike LDA, seqLDA represents documents as a sequence of segments (e.g. sections, paragraphs, or sentences). The seqLDA also restricts itself to the study of sequential topic modeling which is how the sub idea in a segment is associated with its antecedent and subsequent segments. The topic distribution for each segment then depends on its preceding segment [2]. This association is intu-

itively correct given how writing is typically arranged, especially in textbooks where sections and paragraphs should have clear divisions in topics.

Considering our need to categorize sequential sentences together by topic, we chose to use seqLDA over LDA. Given that our final result is expected to be a rough outline of the given textbook section, seqLDA would greatly benefit our goal due to its consideration of the sequential underlying structure.

For sentence classification we needed a classification method which could be trained to identify and extract important features of detail and example type sentences. In the end, we decided to use Convolutional Neural Networks (CNN) due to their characteristics of combining feature extraction and classification while still being general enough to be applicable to various data sets [3]. These characteristics are invaluable when it comes to classifying sentences which could be structured differently between different textbooks. Even in the field of natural language processing, CNN's have shown a remarkable performance on sentence modeling and classification tasks [7].

3. Proposed Technical Plan

3.1. Data Set

For our data set we will be using *Human Geography: People, Place, and Culture 9th Edition* (Erin et al., 2009). We chose this due to this textbook having a reasonably clear distinction between topics. Paragraphs are also organized so that a mixture of topics exist within a single paragraph. Due to time and memory constraints, only select sections of the textbook will be used. The paragraphs also contain many easily differentiable detail and example sentences which can be used to train our CNN. We are also planning on using various Wikipedia paragraphs with clearly defined topics in order to validate whether or not our topic grouping is accurate.

3.2. Pre-processing Textbook Text

Before doing any computations on the text, we plan to first remove stop words. We decided to use the NLTK python library which has dedicated functions to do so. By excluding stop words (such as "a" and "the"), we hope to reduce the noise which can be found in our data set.

3.3. Grouping Sentences via Topic

Once the input text has been preprocessed, we will then split the text into paragraphs and use seqLDA on each paragraph. For our research, we will define segment to be a sentence and documents to be a paragraph. Once the topic is found, we can then group sentences by examining the words found in each topic and determining what percentage of each topic is found in all of the sentences. The topic that occupies a majority of the sentence (greater than 50%) will be categorized as the main topic. Sequential sentences with the same main topic will then be grouped together, forming a sentence group.

3.4. Sentence Classification

Each sentence of the sentence groups will then further be classified as either a main idea, details, or example based on the content that is contained in that sentence. For the scope of our project, the main idea sentence will be the first sentence in the sentence group when the topic is first mentioned. Detail sentences will be defined as sentences which expand on the main topic such as definitions and descriptions. Example sentences will be defined as sentences which describe an application of the topic. Key phrases such as "for example" and "for instance" and a high usage of named entities will also denote an example. Since our goal does not rely on the semantic meaning of words, we will be translating input sentences into their respective parts of speech, named-entities, and only retaining key words and phrases. The CNN will then be given these inputs to be trained on to create an accurate sentence classification model.

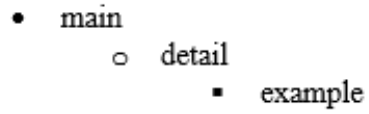

Figure 1. Formatted outline output

3.5. Evaluation

Our goal is to create a rough outline for an inputted section of a textbook in order to create a better reading experience for readers. In order to evaluate whether our research achieves this goal, we will survey students at Texas A&M University using a numeric rating (1 to 5) on how clearly they thought topics were differentiated, how their reading experience was, and whether the outline proved to be more readable in comparison to the original textbook format.

References

[1] David Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. volume 3, pages 601–608,

01 2001.

[2] Lan Du, Wray Buntine, Huidong Jin, and Changyou Chen. Sequential latent dirichlet allocation. *Knowledge and Information Systems - KAIS*, 31, 06 2012.

[3] Lars Hertel, Erhardt Barth, Thomas Kaster, and Thomas Martinetz. Deep convolutional neural networks as generic feature extractors. pages 1–4, 07 2015.

[4] Elizabeth D. Liddy. *Natural Language Processing.* 2001.

[5] Deepak Sharma. A survey on journey of topic modeling techniques from svd to deep learning. *International Journal of Modern Education and Computer Science*, Vol. 9:PP.50–62, 07 2017.

[6] Richard W. Strong, Harvey F. Silver, and Gregory M Perini, Matthew J.and Tuculescu. *Reading for Academic Success: Powerful Strategies for Struggling, Average, and Advanced Readers, Grades 7-12.* 2002.

[7] Rui Zhang, Honglak Lee, and Dragomir Radev. Dependency sensitive convolutional neural networks for modeling sentences and documents, 2016.