Text Book to Lecture Notes using Machine Learning

Mrs. Veena bhat [1] , Nandish kumar A T [2] , Naveen R [3] , Pratiksha krishna jadhav [4], Rachana P [5]

[12345] veena.bhat@amceducation.in , nandishtkumar55@gmail.com, pratikshakj2001@gmail.com ,

prachanar3@gmail.com, naveen.r132002@gmail.com.

[2345]Student , Department of Computer Science and Engineering , AMCEC , Bangalore- 083, Karnataka

[1]Faculty , Department of Computer Science and Engineering , AMCEC , Bangalore- 083, Karnataka

Abstract

As presentations are an essential tool for informing students, making them can take quite a bit of time and work. Using Java-based natural language processing, we propose a novel method in this study for automatically transforming textbooks to presentations. To extract essential concepts and summarise the material, our approach uses named entity recognition, part-of-speech tagging, and dependency parsing. Additionally, we use machine learning algorithms to create clear and cogent slides by identifying connections between these ideas. Our strategy has a number of benefits, including a decrease in the time needed to develop presentations, uniformity in the delivery of the content, and improved student learning. Results from the evaluation illustrate that our approach is accurate and coherent, demonstrating its potential for usage in educational settings. Our suggested system can be useful.

***Keyword –** Machine learning algorithms, NLP, Sentence segmentation, Tokenization, Stemming, Sentence reduction, Sentence scoring, Sentence Position value.*

I. INTRODUCTION

Presentations help teachers successfully communicate information to students and are a crucial aspect of the learning process. High-quality presentation creation may, however, be a labour- and time-intensive procedure. Presentations that are instructive, succinct, and entertaining are now more important than ever due to the rising need for online learning. In this paper, we suggest a Java-based approach to natural language processing for automating the transformation of textbooks into presentations.

To identify the main ideas and summarise the text, our system uses methods including named entity identification, part-of-speech tagging, and dependency parsing. Additionally, we employ machine learning techniques to create clear and short presentations by learni+ng the connections between these ideas. Our suggested strategy has various benefits, including requiring less time and effort to.

II. IMPLEMENTATION OF MODULES

This stage is the underlying stage in moving from issue to the course of action space. Accordingly, starting with what is obliged; diagram takes us to work towards how to full fill those requirements. System plot portrays all the critical data structure, record course of action, yield and genuine modules in the structure and their Specification is picked. This assumes an essential part on the grounds that as it will give the last yield on which it was being working.
In our work we use four modules, these modules are listed below:-

1. Pre-processing of input document:

The phase of pre-processing involves chopping the paragraph into words. This phase involves four stages.
1. Sentence segmentation
2. Tokenization
3. Stop word Removal
4. Stemming

In each stage the document undergoes different changes. The changes are explained below

1.1 Sentence Segmentation of paragraph in the document

Sentence Segmentation is the process of breaking down/segmentation the given text document into sentences al. In this system sentence is segmented by identifying the boundary of sentence which ends with period symbol (.), question mark (?), exclamatory mark (!) and the total number of sentences present in the document are also identified.

1.2 Tokenization of segmented sentences

Tokenization is the process of breaking down the sentences into words. Tokenization is done by identifying the spaces ( ), comma (,) and special symbols between the words. In this process frequency of each word is calculated and stored for further processing.

1.3 Stop Word Removal from the list of words

Stop words are the words that do carry as important meaning as by keywords. These words are identified by supplying a list of words with less importance to the system. The system compares these stop words with the tokenized words obtained from previous phase. These stop words are then disposed as they can interfere and influence the summary that will be generated at the end.

1.4 Stemming

A word can be found in different forms in the same document. These words have to be converted to their root form for simplicity. This process is known as Stemming. An algorithm is used to transform words to their root forms. In this system, Porter's stemmer method is used to turn a word into its root form using a predefined suffix list. Finally, frequency of each is word is calculated a retained for next phase.

2. Sentence scoring

After module 1 the input document is segmented into collection of words in which each word has its individual frequency. In module 2 the sentences are ranked based on seven important features:
1. Frequency
2. Sentence Position
3. Cue words
4. Similarity with the Title.
5. Sentence length.
6. Proper noun.
7. Sentence reduction.

2.1 Frequency

Frequency is the number of times a word occurs in a document. If a word's frequency in a document is high, then it can be said that this word has a significant effect on the content of the document. Salient sentences/words are those sentences/words that occur repeatedly. The frequently occurring word increases the score of sentences they are in. The most common measure widely used to calculate the word frequency is TF (Term frequency) IDF (Inverse document frequency). The total frequency value of a sentence is calculated by summing up the frequency of every word in the document.

2.2 Sentence Position Value

It depends on our requirement whether important sentences are located at certain position in text or in paragraph. Sentences in the beginning define the theme of the document whereas sentences in the end conclude or summarize the document.

The positional value of a sentence is calculated by assigning the highest score value to the first sentence and the last sentence of the document. Second highest score value is assigned to the second sentence from starting and second last sentence of the document. Remaining sentences are assigned a score value of zero.

2.3 Cue Words

Cue words are the important words in a document. These Cue words are given as input from the user. If a sentence contains these Cue words then score value one is assigned to the sentence, otherwise the score value of the sentence will be zero.

2.4 Similarity with the Title

The words in the title and heading of a document that reappear in sentences are directly related to summarization. These words are considered for summarization as the have some extra weight in them. If a sentence contains words in title and header then score value one is assigned to that sentence, otherwise score value is zero for the sentence.

2.5 Sentence length

The length of the sentence resembles the importance of sentence in summarization. Generally, sentences that are very long and very short are not suitable for summary. Sentences that are very long will have unnecessary information which is not useful for summarization of document. Whereas, sentences that are too short do not give much of information about the document.

2.6 Proper Noun

Proper nouns play an important role in summarization. It gives information regarding, to whom or to what the author is referring. Roles played by individuals or locations description will be different more number of times in a document.

2.7 Sentence reduction

Sentence reduction is the method of removing irrelevant phrases like prepositional phrases, clauses, to infinitives, or gerunds from sentences. The goal is to identify less important phrases in a sentence using reduction decisions. The reduction decisions are based on syntactic knowledge, context, and probabilities computed from corpus analysis.

The final score is a Linear Combination of frequency, Sentence positional value, Cue Words, Similarity with the title of the document, Sentence length and Proper noun.

3. Sentence Ranking

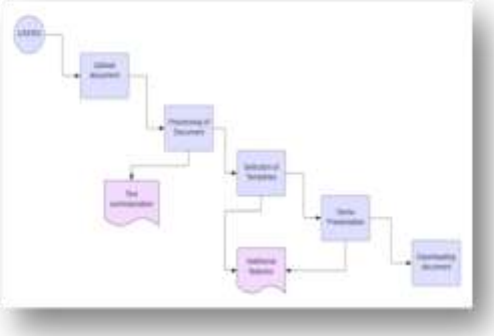
After each sentence is scored they are arranged in descending order of their score value i.e. the sentence whose score value is highest is in top position and the sentence whose score value is lowest is in bottom position

4. Summary Extraction

After ranking the sentences based on their total score the summary is produced selecting certain number of top ranked sentences where the number of sentences required is provided by the user. For the reader’s convenience, the selected sentences in the summary are reordered according to their original positions in the document.

III. ARCHITECTURE DESIGN

An architecture diagram is a visual representation of the structure of your presentation. You can use software like Microsoft Visio or Google Drawings to create your architecture diagram. In the diagram, you should include the main topics or sections of your presentation and how they relate to each other. Mapping Content to Architecture Diagram: Once you have your architecture diagram, you can start mapping the content from the textbook PDF to the diagram. This involves identifying the relevant sections of the textbook and placing them in the appropriate place on the diagram. You can use colour coding or icons to differentiate between different types of content, such as text, images, or diagrams. Creating Slides: After mapping the content to the architecture diagram, you can start creating slides for your presentation. You can use presentation software like Microsoft PowerPoint or Google Slides to create your slides. Each slide should correspond to a section of your architecture diagram and should include relevant text, images, and diagrams. Formatting and Refining: Finally, you can format and refine your presentation to make it more engaging and visually appealing. This may involve adjusting font sizes and styles, adding animations or transitions between slides, and incorporating multimedia elements like videos or audio clips.



IV. CONCLUSION

In conclusion, creating a presentation from a textbook PDF using an architecture diagram is a structured and systematic process that involves several key steps. By selecting a high-quality PDF, creating an architecture diagram, mapping content to the diagram, creating slides, and formatting and refining the presentation, you can effectively communicate the information from the textbook in a visually engaging way. It is important to approach this process with careful planning and attention to detail, and to use appropriate software tools to create the architecture diagram and presentation slides. By following these steps, you can create a high-quality presentation without plagiarism that effectively conveys the information from the textbook to your audience.

V. REFERENCES

[1] NCTD - National Center for Tactile Diagrams. Updated information available at: http://www.nctd.org.uk/ (accessed October 2008).

[2] Enabling Technologies Inc. Updated information available at: http://www.brailler.com/ (accessed January 2010).

[3] NLS - National Library Service for the Blind and Physically Handi capped. Updated information available at: http://www.loc.gov/nls/ (ac cessed January 2010).

[4] R. Velazquez and E. Pissaloux, “Tactile displays in human-machine interaction: four case studies”, The International Journal of Virtual Reality, 7(2), pp 51-58, 2008.

[5] H. Hernandez, E. Preza, and R. Velazquez, “Characterization of a piezoelectric ultrasonic linear motor for Braille displays“, Proc. of IEEE-CERMA, pp 402-407, 2009.

[6] Technologic Systems. Updated information available at: http://www.embeddedarm.com/ (accessed January 2010).

[7] R. Velazquez, E. Preza, and H. Hernandez, “Making eBooks accessible to blind Braille readers“, Proc. of IEEE-HAVE, pp 25-29, 2008.