International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 01 Jan. 2025, pp: 329-334 www.ijaem.net ISSN: 2395-5252

An Overview of Techniques for Concept Extraction and Spaced Repetition in Learning Applications: A Survey for the Development of ”Retain”

Reo George, Rohan Mano, Sourav UP, Gokul Krishna B, Ameena A.

Department of Computer Science College of Engineering, Chengannur.

Date of Submission: 10-01-2025 Date of Acceptance: 20-01-2025

ABSTRACT—The development of educational technologies that enhance learning through automated systems has seen significant growth in recent years. This paper surveys key techniques in concept extraction, spaced repetition, image processing, and automatic text summarization for the creation of a web application named ”Retain.” Retain takes study materials in various formats (PDFs, images) and automatically generates flashcards based on key concepts. The application leverages techniques such as Optical Character Recognition (OCR), spaced repetition algorithms, intelligent content summarization, and multilingual processing to enhance learning efficiency. This survey reviews foundational papers in these areas, offering insights into how these technologies can be integrated to optimize learning outcomes.

Index Terms—Spaced Repetition, Concept Extraction, Optical Character Recognition, Image Processing, Automatic Text Summarization, Multilingual Processing

I. INTRODUCTION

In recent years, educational technology has rapidly evolved, leading to new methodologies for improving student retention and engagement. A key area of innovation has been the development of automated systems that aid in learning by identifying important concepts in study materials and lever- aging spaced repetition to optimize memorization. This survey provides a comprehensive overview of the key research that supports the development of ”Retain,” a web application that processes study materials, extracts key concepts, and creates flash- cards to help users retain information more efficiently.

The scope of this survey includes papers on spaced repetition models, Optical Character Recognition (OCR) systems, advancements in image processing, and automatic text summarization, all of which contribute to the core functionality of Retain. This review explores the challenges, solutions, and future directions in these domains, focusing on their relevance to Retain’s design and performance.

A. Learning Techniques

Spaced Repetition

Spaced repetition has been a cornerstone of educational strategies aimed at combating the forgetting curve, aphenomen on identified by Ebbinghaus which shows that information is lost overtime when there is no attempt to retain it. In spaced repetition, study sessions are strategically spaced to maximize retention and minimize effort.

Memory Models and Adaptive Learning

The efficacy of spaced repetition is grounded in cognitive science, particularly in models like the Power Law of Learning, which posits that the more often amemory is retrieved, the more permanent it becomes. Techniques such as those found in the SM2 algorithm, which is widely used in applications like Anki, take advantage of this by adjusting intervals based on how easily the learner recalls each card.

The Leitner System

TheLeitnersystem,firstintroducedinthe1970s, remains one of the most effective methods of man- aging spaced repetition. In this system, flashcards are sorted into boxes based on how well the learner knows each card. When a card is recalled correctly, it moves to a box with longer intervals between reviews. Cards that are not recalled correctly stay in boxes that are reviewed

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 01 Jan. 2025, pp: 329-334 www.ijaem.net ISSN: 2395-5252

more frequently.

B. Optical Character Recognition

OCR

Optical Character Recognition (OCR) is the technology that converts different types of documents, such as scanned paper documents or images of documents, into editable and searchable data. This is particularly important for Retain, as users may submit study materials in the form of images or scanned PDFs.

Image Captioning and Text Encoders

For study materials that are submitted as images, simply extracting text is not always enough. Often, images contain a combination of text and diagrams that must be interpreted together to generate meaningful flashcards. Image captioning, which involves describing the content of an image, is a key technology in such cases.

In Retain, this technique is used to accurately summarize images that contain important textual content, such as annotated diagrams or info graphics. By intelligently processing both the text and visual elements in an image, Retain can create more accurate and useful flashcards for users.

C. Natural Language Processing

Multilingual Processing

One of the long-term goals for Retain is to sup- port users who study materials in languages other than English. Multi- lingual processing is essential for OCR, text summarization, and spaced repetition algorithms, especially in a globalized education landscape.

Automatic Text Summarization

Text summarization automatically condenses large amounts of text by analyzing it and extracting the most important details into brief summaries. This process is especially vital for applications like Retain, where large volumes of textual data need to be summarized into flashcards that focus on the key ideas.

II. LITERATURE SURVEY

A seminal work in this field is: Su, Jingyong, et al., Optimizing spaced repetition schedule by capturing the dynamics of memory [1], which introduces fundamental theories of spaced repetition, with a particular focus on the Leitner system. The Leitner system, popularized in the 1970s, operates by sorting flashcards into several boxes. Cards that are known well by the user are moved into boxes with longer intervals between reviews, while those that are answered incorrectly are revisited more frequently.

Brendan Meeder's paper, A Trainable Spaced Repetition Model for Language Learning[2],takes this concept further by developing a model that adapts to the learner's individual performance. Unlike fixed-interval models, Meeder's approach trains a system to dynamically adjust review schedules based on each learner's recall accuracy. This method is particularly useful for applications like Retain, where personalization is crucial. By analyzing learner behavior over time, Retain can offer more effective review intervals for each user, thus enhancing learning efficiency.

Adaptive learning algorithms, such as Meeder's model [2], represent the cutting edge of this field. These models assess not only the correctness of a learner's responses but also factors like response time and confidence levels to predict future performance. By dynamically adjusting review intervals, Retain can help learners focus their time on the most challenging concepts, providing a more personalized and effective study experience.

Narkhede, Kshitji, et al. work, Unveiling the Art of Effective Learning through Spaced Repetition and Evidence-Based Techniques [3], outlines the simplicity and effectiveness of this system for optimizing memory retention. The application of the Leitner system in Retain ensures that users are constantly challenged on the concepts they find most difficult, while not wasting time on those they have already mastered. By leveraging this system, Retain is able to balance review efficiency with memory retention, which is especially important for users dealing with large amounts of information.

Jieying Li's research, Researchon English Automatic Recognition System Based on OCR Technology, provides insights into how OCR can be utilized to automatically recognize and convert text from scanned documents into editable text. Modern OCR systems, such as Tesseract, use neural networks to significantly improve the accuracy of text recognition, even in documents with complex formatting or poor image quality [4].

In Retain, OCR plays a critical role in processing study materials. Once text is extracted from an image or PDF, it can then be analyzed for key concepts, summarized, and turned into flashcards. Li's work also discusses methods for improving the speed and accuracy of OCR, which are essential in a real-time application like Retain where users expect immediate feedback.

Arisa Ueda et al.'s paper, Switching Text-

Based Image Encoders for Captioning Images with Text, explores advanced methods for recognizing and interpreting text within images. The ability to switch between different encoding models depending on the content of the image improves the accuracy of captioning, particularly for images with complex layouts or a high density of text [5].

Amirreza Fateh et al.'s work, Advancing Multilingual Handwritten Numeral Recognition with Attention Driven Transfer Learning, discusses the challenges of recognizing text in multiple languages, particularly for handwritten content. Their model employs transfer learning, which allows a system trained on one language to generalize to others with minimal additional training. This is particularly useful in applications like Retain, where users may submit study materials in various languages [6].

The integration of multilingual OCR and summarization tools in Retain would allow the application to serve a wider audience, helping students from different linguistic backgrounds to study more efficiently. By using techniques such as transfer learning and attention-driven models, Retain can be expanded to support a more diverse user base.

Various summarization techniques can be lever- aged to improve the quality and relevance of the generated flashcards. These methods are primarily categorized into two types: extractive summarization, which selects key sentences directly from the original text, and abstractive summarization, which generates new sentences to summarize the content[7].In this section, we explore how different papers approach these techniques and discuss their relevance to Retain.f Bilal Khan, Zohaib Ali Shah, and Muhammad Usman in their paper Exploring the Landscape of Automatic Text Summarization: A Comprehensive Survey [7] provide an extensive review of existing summarization techniques. They classify approaches into extractive and abstractive methods. Extractive summarization works by identifying and selecting sentences or phrases from the original document that best capture the core ideas. This is useful when preserving the original wording is important, which can be particularly helpful for flashcards focused on key definitions or technical terms.

On the other hand, abstractive summarization attempts to generate novel sentences that represent the key information, which makes it more flexible but also computationally more challenging. Retain can benefit from both approaches: extractive summarization for pulling out precise definitions and keywords, and abstractive summarization for generating more concise explanations of concepts.

Furthermore, Khan et al. emphasize the growing role of deep learning techniques, particularly transformer-based models like BERT and GPT, which have improved the quality of abstractive summarization. These advancements can be integrated into Retain to handle large, complex documents, transforming them into useful flashcards without overwhelming the learner.

M.H.H. Wahab etal., in their work A Review on Optimization-Based Automation in Text Summarization [8], delve into various optimization algorithms that enhance the effectiveness and efficiency of automatic text summarization. Wahab and colleagues discus show optimization techniques can be applied to extractive summarization, focusing on selecting the most representative sentences that encapsulate the essence of a document.

These methods are particularly important for Retain, where a balance between summarization accuracy and speed is critical. Users expect flash-cards to be generated quickly, especially when dealing with large volumes of material. Wahab etal. discuss how optimization algorithms such as genetic algorithms and simulated annealing can be applied to finetune the extraction process, ensuring that the generated summaries are both relevant and concise. Implementing such algorithms in Retain could significantly enhance the user experience by speeding up flashcard generation while maintaining high-quality content.

Mridha et al. highlight the challenges associated with automatic text summarization in their paper Survey of Automatic Text Summarization: Progress, Process, and Challenges [9]. They identify several issues such as maintaining the coherence of summaries, handling large-scale data, and dealing with ambiguous or complex language. In Retain, ensuring that the flashcards generated from long or technical documents maintain coherence and usability is essential for effective learning.

The paper also discusses the difficulty in evaluating summarization algorithms, especially for abstractive methods, which do not always align with traditional evaluation metrics like ROUGE (which is commonly used for extractive summarization).For Retain, implementing multiple evaluation strategies, including user feedback mechanisms, could be beneficial to ensure that flashcards are of the highest quality and relevance.

In addition, Mridha et al. suggest that hybrid approaches, combining both extractive and abstractive methods, canmitigate some of these

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 01 Jan. 2025, pp: 329-334 www.ijaem.net ISSN: 2395-5252

challenges. For Retain, this could meanusing extractive methods to identify key phrases and then applying abstractive techniques to condense and explain them, offering a more balanced and comprehensive flashcard generation process.

Yadav et al., in their comprehensive review Automatic Text Summarization Methods: A Comprehensive Review [10], explore the evolution of abstractive summarization, focusing on recent advancement s made possible by neural networks and attention-based models. They particularly emphasize the role of transformer-based architectures, such as BERT and GPT, which have revolutionized the field by improving the model's ability to generate coherent, contextually appropriate summaries.

For Retain, the application of these transformer- based models could greatly improve the quality of abstractive summarization, making it more suit- able for generating flashcards from complex study materials. The ability of these models to generate human-like summaries means that flashcards could be more intuitive and easier to understand for users, offering explanations that are clearer and more concise than purely extractive methods.

Yadav etal. also explore domain-specific summarization, which is particularly relevant for Retain. In fields like medicine, law, or engineering, where specialized vocabulary and complex structures are common, domain-specific models trained on relevant corpora canbe integrated into Retain to ensure that the generated flashcards are tailored to the user's subject area.

III. DISCUSSION AND RESULTS

The literature survey highlights the interplay of various technologies that form the backbone of "Retain," an application designed to revolutionize learning through automation and personalization. Key findings and their implications are summarized below:

Spaced Repetition and Adaptive Learning Spaced repetition techniques, including the Leitner system and adaptive models like SM2, are crucial for enhancing memory retention. The survey demonstrates that adaptive learning algorithms offer a significant advantage by tailoring review schedules to the learner's performance. Retain leverages this adaptability to optimize flashcard review intervals, ensuring a personalized and effective learning experience.

Optical Character Recognition (OCR) and Image Processing

Modern OCR systems, such as Tesseract, enhance Retain's ability to process study materials in di- verse formats, including PDFs and images. These systems handle multilingual content and complex layouts with improved accuracy. Image captioning techniques further extend this functionality by in- terpreting diagrams and annotated visuals, making flashcards more comprehensive.

Text Summarization

A hybrid approach to text summarization—combining extractive and abstractive techniques—ensures that flashcards maintain precision and readability. Extractive methods are effective for identifying key terms, while abstractive techniques generate concise explanations. Transformer-based models like BERT and GPT offer advanced summarization capabilities, enabling Retain to handle complex and lengthy documents effectively.

Multilingual Processing

The survey highlights the growing need for multilingual support in educational applications. Advanced transfer learning models enable OCR and summarization tools to generalize across languages, making Retain accessible to a global audience. Challenges remain in handling domain- specific vocabularies in non-English texts, but emerging techniques are bridging this gap.

User Engagement and Learning Efficiency by integrating these technologies, Retain ensures a focus on challenging concepts while avoiding redundant reviews of familiar material. This balance maximizes user engagement and minimizes cog- native overload, enabling learners to manage large volumes of information more effectively.

Key Results:

- Personalized Learning: Adaptive spaced repetition ensures tailored study schedules, improving long-term retention.
- Efficient Content Processing: Advanced OCR and summarization techniques allow Retain to generate accurate flashcards from diverse materials quickly.
- Enhanced Accessibility: Multilingual capabilities and mobile readiness position Retain as a versatile tool for users worldwide.
- Improved Flashcard Quality: A combination of extractive and abstractive summarization

provides users with concise, relevant, and intuitive content.

By combining these findings, Retain demonstrates the potential to bridge gaps in current educational technologies, offering an innovative platform for students to optimize their learning processes.

IV. FUTURE DIRECTIVES

To enhance Retain's capabilities and address its current limitations, the following five key directives are proposed:

- **Enhanced Adaptive Learning Models** In corporatere inforcement learning or neural based adaptive algorithms to improve personalization further. These models should dynamically adjust not only the review intervals but also the difficulty and presentation style of flashcards based on user engagement, confidence levels, and performance trends.

- **Domain-Specific Summarization Models:** Develop tailored summarization systems for specialized fields like medicine, engineering, and law. These models should focus on extracting and summarizing technical terminology and complex concepts, ensuring the generated flashcards meet the specific needs of learners in those domains.

- **Scalable Real-Time Processing:**

Optimize the infrastructure for real-time processing of large datasets. By improving computation pipelines for OCR and text summarization, Retain can deliver flashcards with minimal latency, ensuring a seamless user experience for high-volume content.

- **Mobile Integration and Offline Functionality:**

Advance multilingual capabilities to include non-Latin scripts and handwritten text. Utilizing transfer learning and fine-tuned OCR models will help handled verse languages and contexts, broadening Retain's accessibility to a global audience.

These future directives focus on making Retain a more versatile, accessible, and user-centric plat- form, setting the stage for its adoption in diverse educational contexts.

V. CONCLUSION

This survey has established a comprehensive framework for the technical development of "Retain," an educational application designed to enhance learning through automation and personalization. By synthesizing state-of-the-art methodologies in spaced repetition, Optical Character Recognition (OCR), text summarization, and multilingual processing, Retain is positioned to significantly improve user learning outcomes.

The review highlights the following critical contributions:

- Adaptive spaced repetition models, such as SM2 and the Leitner system, effectively personalize the learning experience.
- Advanced OCR techniques enable accurate extraction of text and diagrams from study materials, even in diverse formats and languages.
- A hybrid of extractive and abstractive summarization methods ensures high-quality flash- cards that balance detail and conciseness.

While these technologies collectively provide a robust foundation for Retain, challenges such as handling low-quality images, optimizing summarization for large datasets, and scaling multilingual capabilities remain. Addressing these issues will be critical for Retain's success in offering a seamless, effective learning platform.

REFERENCES

[1] Su, Jingyong,etal."Optimizing spaced repetition schedule by capturing the dynamics of memory."IEEE Transactions on Knowledge and Data Engineering 35.10(2023):10085-10097.
[2] Meeder, B. (2021). A Trainable Spaced Repetition Model for Language Learning.
[3] Narkhede, Kshitij, et al. "Unveiling the Art of Effective Learning through Spaced Repetition and Evidence-BasedTechniques."2024 IEEE International Conference on Con-temporary Computing and Communications
[4] Li, J. (2023). Research on English Automatic Recognition System Based on OCR Technology.
[5] Ueda,A., Yang, W.,& Sugiura, K.(2023).Switching Text-Based Image Encoders for Captioning Images with Text.
[6] Fateh,A., Birgani, R.T., Fateh, M., & Abolghasemi, V. (2023). Advancing Multilingual Handwritten Numeral Recognition with Attention Driven Transfer Learning.
[7] Khan, B., Shah, Z. A., & Usman, M. (2023).

Exploring the Landscape of Automatic Text Summarization: A Com-prehensive Survey.

[8] Wahab, M.H.H., Ali,N.H., & Hamid, N.A.W.A.(2022).A Review on Optimization-Based Automation Text Summarization Approach.

[9] M. F. Mridha et al.(2021). Survey of Automatic Text Summarization: Progress, Process and Challenges.

[10] Divakar Yadav, Jalpa Desai, & Kumar Yadav (2021). Au-tomatic Text Summarization Methods: A Comprehensive Review.