ConceptEVA

Concept-Based Interactive Exploration and Customization of Document Summaries

Zhang, Xiaoyu; Li, Jianping; Chi, Po Wei; Chandrasegaran, Senthil; Ma, Kwan Liu

DOI
10.1145/3544548.3581260

Publication date
2023

Document Version
Final published version

Published in
CHI 2023 - Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems

ConceptEVA: Concept-Based Interactive Exploration and Customization of Document Summaries

Xiaoyu Zhang
xybzhang@ucdavis.edu
Department of Computer Science, University of California, Davis
Davis, California, USA

Jianping Kelvin Li
jpkelvinli@gmail.com
Databricks (Work Done at UC Davis)
San Jose, California, USA

Po-Wei Chi
pwchi@ucdavis.edu
(Work Done at UC Davis)
Austin, Texas, USA

Senthil Chandrasegaran
r.s.k.chandrasegaran@tudelft.nl
Faculty of Industrial Design Engineering, TU Delft
Delft, The Netherlands

Kwan-Liu Ma
klma@ucdavis.edu
Department of Computer Science, University of California, Davis
Davis, California, USA

ABSTRACT

With the most advanced natural language processing and artificial intelligence approaches, effective summarization of long and multi-topic documents—such as academic papers—for readers from different domains still remains a challenge. To address this, we introduce ConceptEVA, a mixed-initiative approach to generate, evaluate, and customize summaries for long and multi-topic documents. ConceptEVA incorporates a custom multi-task longformer encoder decoder to summarize longer documents. Interactive visualizations of document concepts as a network reflecting both semantic relatedness and co-occurrence help users focus on concepts of interest. The user can select these concepts and automatically update the summary to emphasize them. We present two iterations of ConceptEVA evaluated through an expert review and a within-subjects study. We find that participants' satisfaction with customized summaries through ConceptEVA is higher than their own manually-generated summary, while incorporating critique into the summaries proved challenging. Based on our findings, we make recommendations for designing summarization systems incorporating mixed-initiative interactions.

CCS CONCEPTS

• **Human-centered computing** → **Information visualization**;
• **Information systems** → *Ontologies*.

KEYWORDS

Interactive Visual Analytics, Document Summarization, Knowledge Graph, Mixed-Initiative Interfaces

ACM Reference Format:
Xiaoyu Zhang, Jianping Kelvin Li, Po-Wei Chi, Senthil Chandrasegaran, and Kwan-Liu Ma. 2023. ConceptEVA: Concept-Based Interactive Exploration and Customization of Document Summaries. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2023), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3544548.3581260

CHI 2023, April 23-28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581260

1 INTRODUCTION

The notion of automated text summarization—compression of long text passages into shorter text without losing essential information—has been an open problem since over half a century ago [35]. The main goals of automated text summarization are to present the salient concepts of a given document in a compact way, and to minimize repetition of the presented ideas or concepts [16]. Earlier techniques fall under the umbrella of extractive summarization where summaries are generated by extracting terms, phrases, or entire sentences from the source text using statistical techniques [17]. With advances in machine learning and specifically sequence-to-sequence language models, abstractive summarization—an approach that generates paraphrased text that still retains concepts from the original text—has gained recent popularity as it mimics summaries created by humans [51].

However, significant challenges in abstractive summarization remain, such as the summarization of long, complex, documents that span multiple knowledge domains. While approaches have been proposed for summarizing domain-specific text [33] and others for summarizing long documents [60], the challenge remains that there is no one "ideal" summary for such long and multi-domain documents. Automated summarization systems typically do not fare well when the source document spans multiple topics regardless of approach, i.e., extractive [18], abstractive, or hybrid [16].

Academic papers, especially those in the fields of design or human-computer interaction (HCI) where research tends to be cross-disciplinary, tend to fall under this category of long, multi-topic documents. For instance, a research article might span the fields of wearable technologies, privacy, and social justice. A summary of this article that is deemed useful by a researcher in wearable technology would be different from one deemed useful to a researcher in security and privacy. Yet, both summaries may still be perfectly valid summaries of the article. This subjectivity means that purely automated, black-box approaches to summary generation will not work. Instead, a human-in-the-loop approach is needed to allow the user to steer the automated summary generator

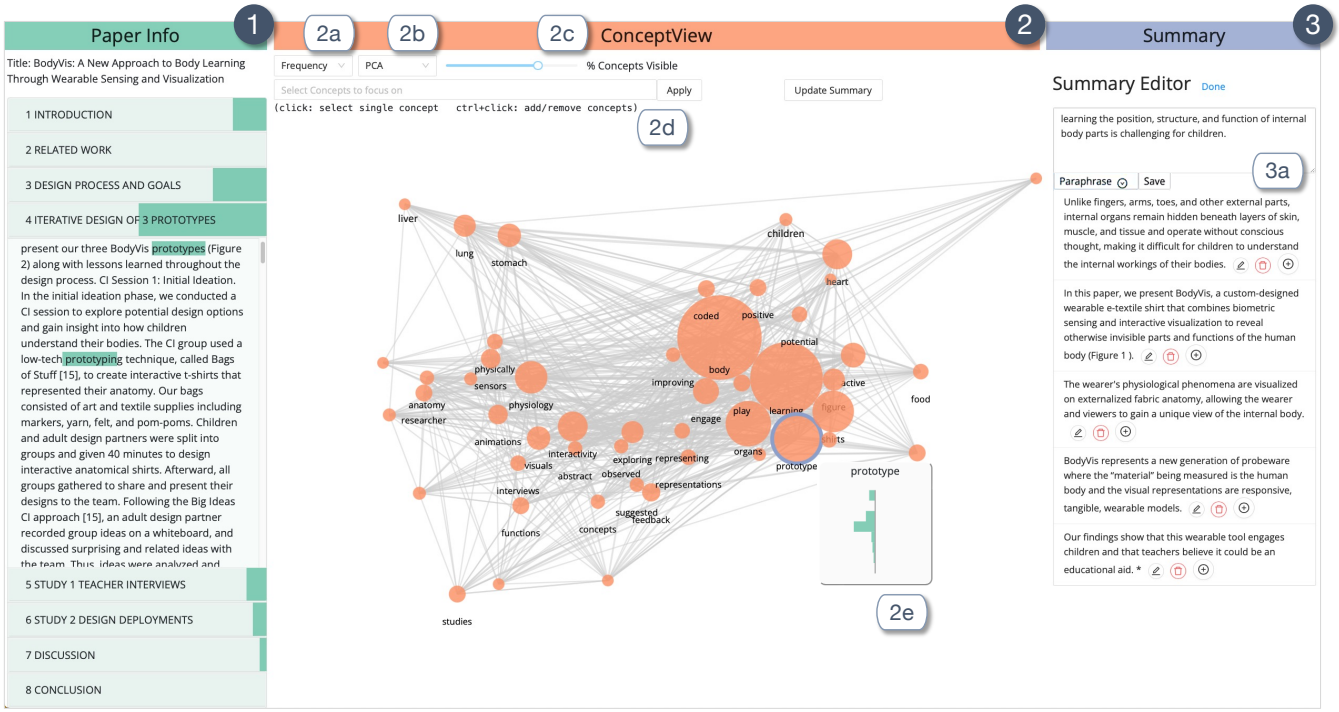


Figure 1: The ConceptEVA Interface shows a multi-disciplinary research paper [43] and its auto-generated summary. The interface can be separated into three main panels vertically. The panel on the left (1) shows a section-wise collapsed view of the research paper, while the panel on the right (3) shows the generated summary in the summary editor. At the center is the concept view (2), which displays the concepts extracted from the paper based on a reference ontology. The circles represent concepts whose layouts are decided by their text embedding and co-occurrence relationship with other concepts. The links indicate whether and how frequently the concepts they connect co-occur in the same sentence. A *concept glyph* comparing the concept's appearance distribution across the document and the summary is shown in (2e) upon user request. Users can also adjust the concept layout in (2) according to their interests by changing (2a) the information encoded with circle size, (2b) the text embedding projection algorithm, (2c) the percentage of concepts visible, or focusing on the concepts they are interested in (2d). They can also interactively edit (3a), reorder (with drag and drop), or delete any sentences in (3). The user can select concepts of interest for the summary generator, which then generates a new summary incorporating the selected concepts.

to interactively generate a summary that is relevant to the user's interests.

To address this challenge, we present ConceptEVA, a mixed-initiative system for academic document readers and writers to generate, evaluate, and customize automated summaries. We build a multi-task Longformer Encoder Decoder (LED) [3] from a pre-trained LED trained for scientific document summarization by fine-tuning it on two downstream tasks—paraphrasing and semantic sentence embedding—to handle long documents. This approach uses the notion of attention mechanisms from transformers [57] at local levels to reduce memory usage, and at global levels to preserve information fidelity in longer documents.

In addition, ConceptEVA also supports summary customization for the user by visualizing the concepts—in this scenario, topics explicitly defined in an ontology or knowledge graphs—occurring in the document. The concepts are identified using a multi-domain ontology [2], and visualized as a force-directed layout of a graph network using metrics such as concept relatedness and concept co-occurrence in the document. We introduce a function that we call "focus-on", that allows the user to select concept(s) of interest to surface and highlight other concepts related to the selected ones. The user can identify concepts to focus on and use them to steer the automated summarizer to generate a summary text in which the concepts of interest feature prominently. The user can also further edit the summary at the concept level by navigating the original document and selecting text to emphasize using the concepts of interest as a filter. They can also edit it at the sentence level by selecting alternative paraphrasing and sentence ordering.

The design of ConceptEVA informed by an initial survey of eight research practitioners, and refined through two stages of development and evaluation:

Iteration 1. A hierarchical summarization approach with a glyph-based visualization of concepts embedded in a two-dimensional projection, evaluated through an expert review of 3 participants;
Iteration 2. The final LED approach to summarization described above with concepts visualized as a force-directed layout that preserves both semantic relatedness and concept co-occurrence

within the document. This version is evaluated using a within-subjects study of 12 participants using manually-generated summaries as a baseline.

Findings from the user study indicate that ConceptEVA is seen as helpful for participants in examining and verifying ideas, and using specific concepts of interest to explore related concepts and how they are addressed in the source document. ConceptEVA was also reported as more useful when the participants evaluated and customized a summary of a document that lay outside their domain of interest, while it was seen as less useful when the participant was knowledgeable about the domain or had a specific idea of what the summary should include. Using ConceptEVA for summarization allowed participants to address content-specific aspects of the summary, but inexeperienced participants found it more difficult to incorporate critique such as limitations and implications into the summary.

The chief contribution of this work is ConceptEVA, a mixed-initiative system that integrates interactive visual analysis and NLP techniques for evaluating and customizing long document summaries. Specifically, we fine tune an LED trained for scientific document summarization for paraphrasing and semantic sentence embedding, identify and visualize concepts from a given academic document using a reference ontology, and provide an interactive visualization system to identify concepts of interest and use them to customize the summary. We also present insights from a user study on how well users are able to follow summarization guidelines when using ConceptEVA. Finally, we make recommendations for future development and analysis of mixed-initiative summarization systems such as maintaining the user's mental map of the original document by preserving its layout, allowing users to create custom groupings of concepts that will help them add critique to the summary, and minimizing interactive latency for a more fluid interface.

2 RELATED WORK

ConceptEVA introduces a human-in-the loop, mixed-initiative approach to evaluate and customize document summary generation. In this section, we review prior work in the domains of summary generation, summary evaluation, and text and document visualization on which we build to create ConceptEVA.

2.1 Summary Evaluation

Summary evaluation techniques can be divided into two main categories: intrinsic [19] and extrinsic [41]. Intrinsic evaluation methods evaluate a summary based on how well its information matches the information in a reference summary, which is typically human-generated. Some examples of intrinsic evaluation of summarization include ROUGE [32] and BERTScore [63]. Bommasani and Cardie [6] propose separate intrinsic scores for compression, topic similarity, abstractivity, redundancy, and semantic coherence. In contrast, extrinsic evaluation methods evaluate summaries based on their suitability to specific tasks such as following instructions, assessing topic relevance, or answering questions [13, 22, 41]. In extrinsic approaches, humans subjects are asked to use different summaries to perform a task and uses metrics for their performance—such as completion time and success rate—to evaluate the summaries.

Our work incorporates the principles behind extrinsic summary evaluation methods. By effectively revealing and comparing the important concepts in a document and its summary, readers can gain confidence in a qualified summary by confirming that it includes all the interested concepts, or see which concepts are missing in a "poor" summary.

2.2 Summary Generation and Customization

Advances in deep learning and AI has made the automatic generation of good-quality summaries for long document text possible, featured by the success of Transformers [57] with its innovative architecture and attention mechanism. Unsupervised pre-training methods—Masked LM (MLM) and Next Sentence Prediction (NSP)—proposed by Devlin et al. [12] for their Bidirectional Encoder Representations from Transformers (BERT) enables modeling natural language on a huge corpus, and then fine tuning the model on downstream tasks like summarization. Inspired by BERT, other researchers [29, 45, 46, 62] propose different pre-training methods and improve the quality of summarization. For instance, Li et al. [30] propose a multi-task training framework for text summarization that trains a binary classifier to identify sentence keywords that guides summary generation by mixing encoded sentence and keyword signal using dual attention and co-selective gates. Wu et al. [60] use a top-down approach to recursively summarize long articles like books. In our work, we use the Longformer Encoder Decoder (LED) [3] for long scientific document summarization, which turns a full attention mechanism—computing relationships between every pair of words in the document—to a local attention mechanism—computing relationships between a more "local window" of limited words that precede and succeed any given word. This has two benefits: faster computation and lower memory usage, which makes it more capable of processing longer documents without a significant drop in the summary quality.

For the summary customization task, most existing NLP techniques utilizes memory to adjust the auto-regressive language model's output distribution such that the models can retrieve external information given the input prompt. Nearest-Neighbour Language Models [25] merge the retrieved information into the output distribution and boost up the language model's perplexity without training. Borgeaud et al. [7] show that by incorporating a large-scale explicit memory bank, a smaller language model can achieve performance comparable to models like GPT-3 with 25 times more parameters, and can update its memory bank without additional training. Inspired by these methods, we apply Faiss [24] to retrieve the k-nearest sentences for each sentence relevant to a chosen concept, and we customize summaries given these sentences as context.

Besides fully automated approaches, there are also semi-automatic solutions that incorporate humans in the loop. Post-editing [28, 39] is a common semi-automatic approach for summarizing text, which allows humans to edit AI-generated summaries to ensure accurate

and high-quality summarization. Compared to post-editing, ConceptEVA's approach better exploits human-AI collaboration and iteratively improves the summary by leveraging such collaboration. In contrast to post-editing which only allow human to edit the summary at the end, ConceptEVAsupports users to iteratively evaluate and refine the summary by inputting their intention on what should be summarized to the AI models. In ConceptEVA's workflow, users can also edit the AI-generated summary. But instead of direct manual editing, ConceptEVAleverages AI models to provide aids, such as connections to the concepts, and suggestions for paraphrasing.

2.3 Interactive Visual Analysis for Text Data

Our work involves designing interactive visualizations of word embedding and thematic infographics to facilitate summary evaluation and customization. Visualization of word embeddings [21, 34, 52] has been used for supporting text data analysis, such as selecting synonyms, relating concepts, and predicting contexts. In a different way, thematic visualizations are useful for exploring document and conversational texts. For instance, ConToVi [14] uses a dust-and-magnet metaphor [53] to visualize the placement of conversational turns (dust) in relation to a set of topics (magnets). NEREx [15] provides a thematic visualization of multi-party conversations by extracting and categorizing named entities from transcripts. The conversation is then visualized as connected nodes in a network diagram, allowing a visual, thematic exploration of the conversation. TalkTraces [9] uses a combination of topic modeling and word embeddings to visualize a meeting's conversation turns in real time against a planned agenda and the topics discussed in prior meeting(s). VizByWiki [31] automatically links contextually relevant data visualizations retrieved from the internet to enrich new articles. Kim et al. [26] introduced an interactive document reader that automatically references to corresponding tables and/or table cells. All these works exploited visualizations to provide contexts or additional information for helping readers to better comprehend text contents.

The application of concept-based clustering is not limited to text analysis: Park et al. [44] cluster neurons in deep neural networks based on the concepts they detect in images, and in addition create a vector space that embeds neurons that detect co-occurring concepts in close proximity to each other. Berger et al. [4] propose cite2vec, a visual exploration of document collections using a visualization approach that groups documents based on the context in which they are cited in other documents, creating a combined document and word embedding. Closest to our own proposed work is VitaLITy [42], an interactive system that aids academic literature review by providing a mechanism for serendipitously discovering literature related to a topic or article of interest. VitaLITy uses a specialized transformer model [11] to aid academic literature recommendations that use additional data such as citations. These recommendations are presented via a 2-D projection of the document collection embeddings generated from the transformer model. Our work also uses word embeddings to project a view of relevant concepts onto a 2D space, but is different from VitaLITy in the purpose: our focus is on interactively exploring the concept focus of a generated summary as well as generating summaries that emphasize concepts of interest within an academic publication.

In our work, we use visualization of word embeddings to provide overviews of all the important concepts in a document and identify which concepts are missing in the summary for evaluation. Thematic infographics is used in the visualization of word embedding to show the details and occurrences of a concept in both the document and summary for comparison.

3 DESIGN REQUIREMENTS

In order to better understand the different requirements and motivations when summarizing an academic article, we conducted a preliminary survey of 8 higher education professionals: one professor, 4 associate professors, and 3 assistant professors (7 male, 1 female, all between 25–44 years of age). The survey covered open-ended questions concerning how they motivated and guided students' paper summaries, how they evaluated such summaries, and what they consider to be a good summary and why.

Based on the experts' responses, we grouped their remarks and suggestions under three categories: ***process***, representing approaches they use or suggest students to follow in order to summarize an academic document; ***content***, representing what should be included in the summary; ***requirements***, representing attributes that make for a "good" summary. Each remark or statement below is suffixed with a count showing the number of experts who shared the corresponding opinion.

- Process: Approaches to follow when summarizing.
  - Prioritize referring to the abstract, conclusion, introduction, and title (7 experts).
  - Use the abstract & introduction as a "backbone" for the summary (1 expert).
  - Familiarize oneself with background and context, then identify strengths & weaknesses (1 expert).
  - Find parts of the paper relevant to one's context or interest and focus on them (1 expert).
- Content: What the summary should include.
  - An Explanation of what the paper is about and what its contributions are (5 experts).
  - The major ideas of the proposed solution and its difference from prior work (3 experts).
  - The results generated by the solution, and how they address the problem/research question (3 experts).
  - The problem addressed by the paper and the research questions it answers (2 experts).
  - An outline of existing approaches to address the research question or problem, their advantages and limitations, and the challenges (2 experts).
  - The advantages/disadvantages of the solution and the strengths/weaknesses of the paper (2 experts).
- Requirements
  - The summary should have an indication that the summarizer has not simply paraphrased the paper but also thought about and understood the underlying ideas (3 experts).
  - The summary should show reflection on the ideas and discuss implications for practice/research. (3 experts)
  - The summary should include a figure if possible (2 experts).

- The summary should have a clear structure & emphasis (2 experts)
- The summary should be specific and provide details, paraphrasing where necessary and quoting from the paper where necessary (1 expert).

While the above responses are relevant for manual summarization, we also examined existing approaches of evaluating automated summarization techniques, such as fluency, saliency, novelty, and coherence [54]. Saliency is an especially complex issue as saliency of a given summary may vary across readers depending on each reader's background and research focus. Based on the responses and on prior work on automated summarization, we synthesized the following requirements that we prioritize for mixed-initiative approaches that help the user evaluate and customize summaries of scientific articles:

R1 Accuracy Evaluation: The technique should help the user efficiently verify whether a summary accurately reflects the content of the original document based on the criteria established by the user (see R4: Flexibility below). This requirement is synthesized from participant responses categorized under *"criteria"* and *"structure"*.

R2 Provenance Evaluation: The technique should show direct or indirect contributors to a summary to help the user verify whether the summary reflects the structure and key components of the original document. This includes the parts of the original document—a research article in this case—that contribute to the summary. It also includes external references (see R3: Contextualizations) that influence parts of the summary. This requirement is synthesized from responses under *"topics"*, *"structure"*, and *"strategies"*.

R3 Contextualization: The technique should be able to provide some context in which the work presented in the paper exits. Such a context includes the contribution of the work, as well as the significance of the work, its strengths, weaknesses and so on. This can include information presented within the paper itself but should not be restricted to it. This requirement is based on the participant responses under *"criteria"*.

R4 Flexibility: The technique should be flexible enough to change the summaries based on the priority of the user. For instance, the summary may focus on the relevance of the paper to a concept of interest to the user. Alternatively, the summary may also be one that examines the paper's contributions, approach, and methods—or any combination thereof. The requirement is based on participant responses under *"topics"* and *"strategies'*.

4 METHODOLOGY

In ConceptEVA, we support summary evaluation and customization by empowering the exploratory visual analysis (EVA) with multiple natural language processing (NLP) techniques. In this section, we first introduce the data processing and visual analysis framework of ConceptEVA, then describe the major NLP techniques backing the functionalities.

4.1 Framework Overview

ConceptEVA leverages knowledge graphs, NLP, and EVA techniques to facilitate summary evaluation and customization for academic document readers. We bridge the original document and the summary with a concept view visualizing all of the concepts identified from the document. As shown in Fig. 2, we start by extracting concepts from an academic document according to a reference ontology, converting them into text embeddings and projecting them onto a two-dimensional space (Sec. 4.1.1). After that, we present the semantic and contextual information of the concepts in an interactive visual interface that supports flexible concept exploration and customized concept(s) prioritizing (Sec. 4.1.2). Finally, we provide an interactive summary editor to facilitate dedicated refinement of a new version of the summary we generated according to the user-specified concepts of interest (Sec. 4.1.3). In this way, we help the users evaluate the quality of an AI-generated summary and see how well it addresses the readers' focus of interest in the paper, as well as support them customizing the summary to alter their specific requirements if the automated one is not satisfied enough.

4.1.1 Concept Extraction and Projection. In order to effectively extract the key concepts from a large body of texts, knowledge graphs, such as DBpedia [2], Freebase [5], and Wikitology [47], can be used to look up established concepts in specific domains. We use DBpedia-Spotlight [37] to extract concepts and rank their importance by term frequency. We then visually highlight concepts to show which ones are included or missed in the AI-generated or customized summary. To vectorize these concepts, ConceptEVA leverages text embeddings to represent concepts, sentences, and descriptions of the concepts as high-dimensional vectors. Two-dimensional projections of these "concept vectors" are computed using dimensionality reduction techniques, such as PCA[55], t-SNE [56], or UMAP[36]. Semantically similar concepts are placed closer together in the projections, while different concepts are placed farther apart.

4.1.2 Exploratory Visual Analysis. To allow readers to explore and reason about the concepts, ConceptEVA provides interactive visualizations to help trace these concepts back to the source document text as well as to the generated summary. A visual representation (see Sec. 5 for details) is designed to show the importance of the concepts and help the user compare their occurrences in the document text and the summary. Readers can not only use ConceptEVA's interactive visual interface to explore and understand each concept, but also select concepts that are relevant to their research interests. The selected concepts are used to recompute the importance and relevance of each concept in the high-dimensional embedding and recreate the projection, allowing the readers to "steer" the exploration.

4.1.3 Summary Editing with Mixed-Initiative Interactions. While generating a good summary that can satisfy the user's needs and interests cannot solely rely on NLP techniques, ConceptEVA provides a set of mixed-initiative interactions for quickly customizing and editing an AI-generated summary. From the user interface, users can easily select which concepts in the document are important or match their interests. If the generated summary did not provide enough context or description of these concepts, the user can indicate where in the summary that they want to add a sentence about a particular concept, then ConceptEVA will immediately generate a list of sentences that describe that concept for the user to choose.

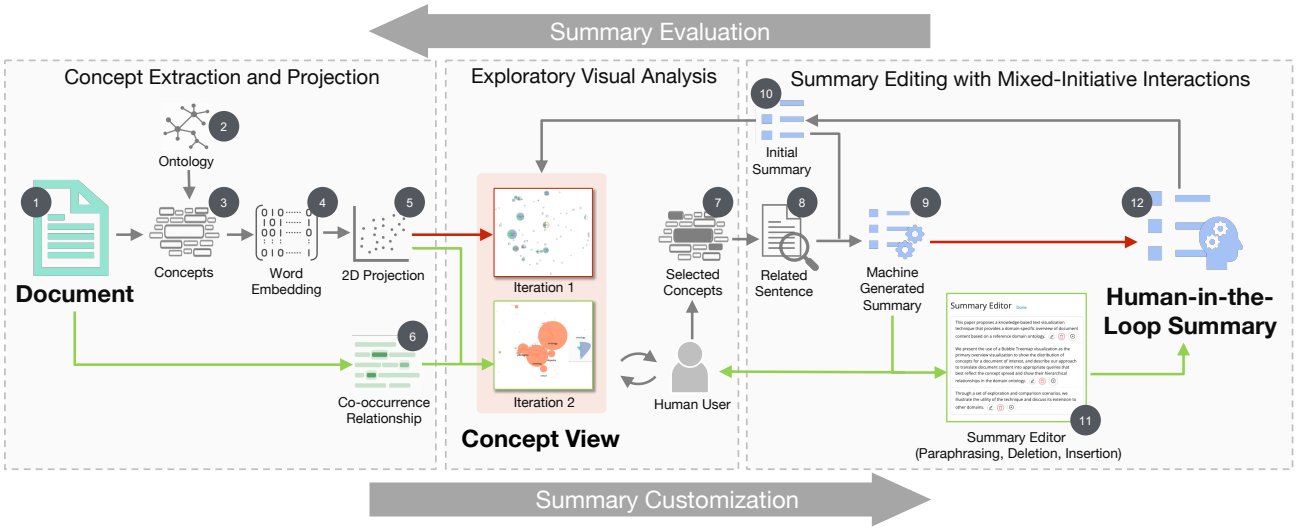


Figure 2: The framework of ConceptEVA. The core idea is to bridge the document and the summary with a concept view. In iteration 1, the concept view shows an embedding-based layout that allows users to select concepts to include in the machine-generated customized summary (red boxes & arrows). In iteration 2, the concept view also includes the co-occurrence information in a force-directed layout, and a summary editor with mixed-initiative interactions is added (green boxes & arrows). In both iterations, the user can repeat the human-in-the-loop summary customization for multiple rounds till they are satisfied with the result.

In addition, ConceptEVA allows users to paraphrase any of the sentences based on its NLP models.

4.2 Natural Language Processing: Multi-Task Longformer Encoder Decoder

As shown in Fig. 2, ConceptEVA uses several NLP techniques at various stages of summary generation and customization. At the center of these techniques is a multi-task Longformer Encoder Decoder (LED) [3] that we develop for iteration 2. We describe in this section the motivation to use LED and its functions at specific stages in summary generation and customization.

In the first iteration of ConceptEVA, we developed a a hierarchical summarization method with BERT Extractive Summarizer [38] and a Pegasus abstractive summarizer [62] for summary generation and customization of long documents (please refer to supplementary materials for more details). However, this approach could easily incur high interaction latency caused by sentence clustering and iterative summarization of long documents. To alleviate these issues, we develop for the second iteration a multi-task Longformer Encoder Decoder (LED) [3], capable of processing longer documents. In addition, we take advantage of weight sharing, i.e., every task shares weights on the common parts of the network's memory, thus optimizing the time and space efficiency of ConceptEVA and speeding up the system's responses to human input.

Our multi-task LED is employed in ConceptEVA for four functionalities: scientific document summarization, paraphrasing, semantic text encoding, and summary customization (see Fig. 3). We describe these functionalities below.

Scientific Document Summarization: The LED model was trained on the ArXiv dataset of scientific papers [10]. Due to its local self-attention mechanism, the memory complexity of LED grows linearly, making it capable of handling up to 16384 tokens, which is typically long enough for handling academic papers. These factors render the LED suitable for generating summaries of academic papers. These automatically-generated summaries (see item '10' in Fig. 2) act as a starting point for users to evaluate and customize upon according to their interests.

Text Paraphrasing: One of the functions in ConceptEVA's mixed-initiative interactions is the ability to paraphrase text, or specifically, generate alternative summaries for a selected sentence. To achieve this capability, we fine tune the pre-trained model on relatively small datasets with small learning rates. We "freeze all the layers" of the model, i.e., we keep all model weights the same during training except for the last two decoder layers. The decoder takes a sequence of tokens as the input and generates the next token based on its weights. We train these two decoder layers on a dataset that contains 147,883 sentence pairs, with each pair containing two alternative paraphrases of one sentence (Fig. 3b). We build this dataset by merging three other datasets: PAWS [65], MRPC from GLUE [58], and TaPaCo [50]. Once fine-tuned, this model is capable of taking as input one sentence and providing a paraphrased sentence as an output. In item '11' in Fig. 2, this model is accessed via the summary editor when the user opts for automated paraphrasing of a selected sentence.

Text Embedding: To generate the concept layout (see Fig. 1-2) and fetch relevant context for summary customization, text embeddings—representing the relationships between concepts or

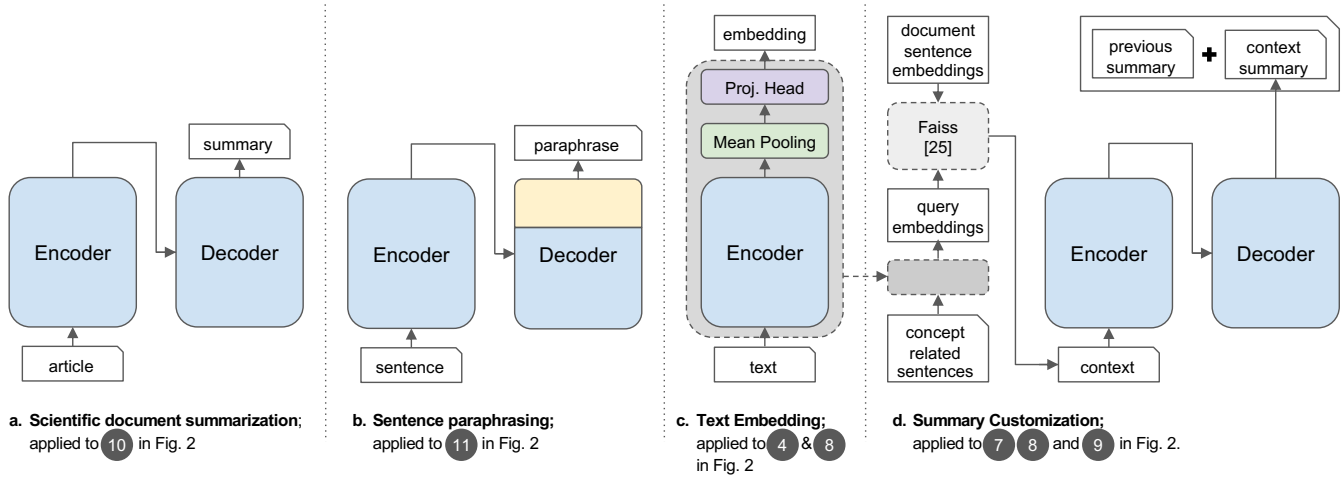


Figure 3: ConceptEVA uses a multi-task LED model [3] to help generate, evaluate, and customize summaries. Specifically, LED performs four functions, shown above as subfigures a–d with detailed explanations in Sec. 4.2. The text below each subfigure indicates the corresponding function in Fig. 2 for which the model is used. In each subfigure, the blue rounded boxes represent the weights from the LED trained for summarizing scientific papers and shared across all tasks. The yellow, purple, and green rounded boxes represent fine-tuned layers for downstream tasks. The functions are: (a) Scientific document summarization: The LED's training data, local self-attention mechanism, and high memory complexity make it suitable to summarize academic papers. (b) Sentence paraphrasing: We fine tune the last two decoder layers (shown in yellow) with a set of "paraphrasing datasets"—datasets that contain multiple paraphrases of a given set of sentences. This helps in generating alternative sentences for a given sentence when editing a summary. (c) Text embedding: To generate the concept layout (see Fig. 1-2) and fetch relavant context for summary customization, we compute text embeddings—vector representations of concepts or sentences in a high-dimensional space. This is done by adding a mean pooling layer (green) and a projection head (purple) to the encoder and fine-tuning it (see Sec. 4.2 for details). (d) Summary customization. Pre-computed embeddings of every sentence in the source document are queried using vector representations—retrieved from the text embedding shown in (c)—of user-selected concepts. Nearest sentences are appended to provide 'context' for the selected concepts, and then summarised. The resulting summarized sentences are appended to the existing summary (see details in Sec. 4.2).

sentences in a high-dimensional space—need to be computed. To compute sentence embeddings, we follow the siamese network architecture from SentenceBERT [49], an approach to generate sentence embeddings, i.e., vector representations of sentences that preserve semantic relationships. We add a 'mean pooling layer'—a function that averages the embeddings of input tokens—and a 'projection head'—a function that computes a high-dimensional space that captures semantic similarities between all sentences—on the LED's encoder (Fig. 3c). We then fine-tune the encoder for learning meaningful sentence embeddings by freezing all layers of the encoder and only training on the projection head. For the training data, we once again follow SentenceBERT: we combine the SNLI [8] and MultiNLI [59] datasets, and format each data sample as a triplet of an 'anchor sentence', a 'positive sentence', and a 'negative sentence'. The training involves fine-tuning the embedding such that in each triplet, the positive sentence ends up closer to the anchor sentence than the negative sentence. We also follow data augmentation approaches (detailed in the supplementary materials) inspired by those followed in SentenceBERT [49]. The resulting model is used in two main functions of ConceptEVA: generation of the "concept view" (see Fig. 2), the "focus-on" function (detailed in Sec. 5.2), and subsequent summary customization (see items '7' and '8' in Fig. 2).

Summary Customization: ConceptEVA customizes a generated summary by updating it to include concepts of interest selected by the user. To achieve this, we pre-compute embeddings for every sentence in the source document. When a user selects a concept or concepts of interest, we retrieve corresponding text embeddings using the model described in the previous paragraph. We then use these embeddings as 'queries' to search for sentences in the pre-computed embeddings that are closest to the query vectors (see Fig. 3). We apply Faiss [24]—a similarity search library of dense vectors in large scale—to implement this approach. The nearest sentences are concatenated in the order of their appearance in the original document and included in the input to the summarizer as 'context' for the selected concepts. The resulting, newly-summarized sentences are then appended into the previously-generated summary. In this form of summary customization, new concepts add to the existing summary but do not result in the erasure of parts of the existing summmary. The summary editor provides the option for the user to manually delete the sentences.

5 INTERFACE DESIGN

The ConceptEVA interface (Fig. 1) consists of three main panels: a document view on the left (with green header & accents) that collapses into a section-wise overview, a summary view (blue header

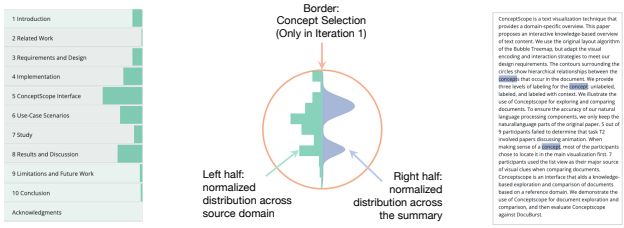


Figure 4: The *concept glyph* extends the concept circle to support the in-place comparison of concept distribution between the document and the summary. This glyph is shown for all dominant concepts in iteration 1 and in a floating tooltip upon request in iteration 2.

& accents) on the right displaying the generated summary and associated metadata, and a central concept view (orange header & accents) showing the relative dominance and associations between the concepts found in the document. Additional controls for visualizing and filtering the concepts are also provided on top of the concept view. The interface design has gone through two iterations of development, incorporating feedback and insights from the expert review (Sec. 6). We detail the visualization and interaction design choices of the final version of the system and the underlying rationale in this section.

5.1 Concept View: Document-Summary Relations

In the concept view, we provide an overview of the document-summary relation from the perspective of concepts. We represent each of the concepts occurring in the documents as a node—a "concept circle"—the size of which shows the dominance of the concept in the source document. User-specific metrics of dominance, such as "frequency" and "tf-idf" are available for the user to choose.

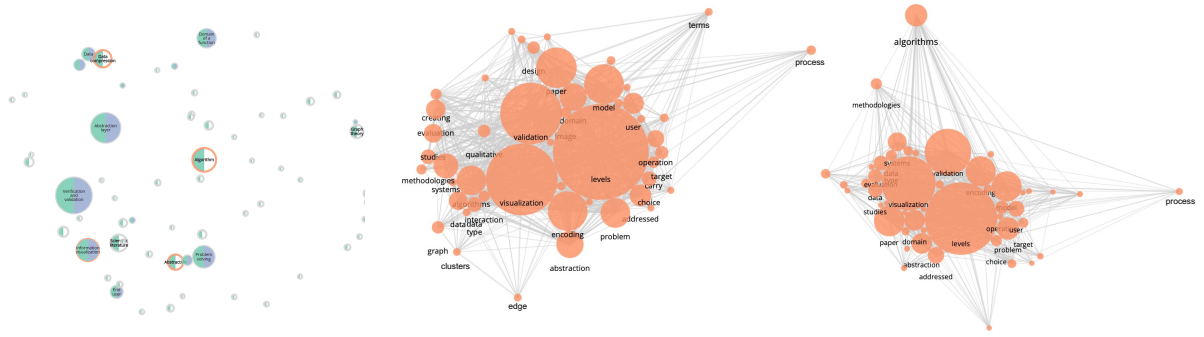
To convey information about the structure of the document and of the summary (**R2**), we incorporate the user's orientation to the interface—the document on the left and summary on the right—into the concept view to represent concepts that are present in the document and concepts present in both the document and the summary. We design the *concept glyph*—a pair of histograms representing the distribution of the concept across the source document and the summary respectively (see Fig. 4). The histograms are oriented vertically and share a common axis. This way, the histogram on the left indicates the source document and the curved line on the right (histogram smoothed with a kernel density estimation) represents the summary. The number of bins on the histogram on the left matches the number of sections in the source document, while the right one maps to the number of sentences in the summary. For instance, the concept "prototype" is missing in the summary shown in Fig. 1 because the right half of the glyph is missing. To further reinforce this connection between the histogram and the document view, we create an echo of the histogram overlaid on top of the section headers (Fig. 4-a). This allows the user to identify the sections of the document in which the concept is most dominant, and examine those sections closely if needed.

When determining the two-dimensional(2D) layout of these concepts on the concept view and the amount of information to reveal for each of them, we started with an embedding-based layout in iteration 1 where the *concept glyph* of every concept were displayed and distributed according to the text embedding (Fig. 5-a, see supplementary materials for details.) While this layout was designed to help the user efficiently compare the occurrence of concepts in the original document against those in the summary, the expert review results (Sec. 6) indicated that showing such a comparison for all the concepts in one visualization was too overwhelming to the users. To reduce such perception load, we shifted to a more intuitive visualization design in iteration 2 where the visual representation of the concepts were simplified to solid circles (Fig. 5-b) in a force-directed layout. The coordinates of these circles are initialized by a 2D projection of the concepts' semantic word embedding and adjusted by links representing the co-occurrence relationship of two concepts in the same sentence per experts' request for more co-occurrence information support. In this way, we created a context-augmented layout with the coordinates of each concept influenced by both its semantic meaning and its co-occurrence relationship with the other concepts in the specific academic document (**R3**). For instance, Fig. 1-2 shows that the concepts "organ" and "prototype" are semantically remote but co-occur frequently in [43], while aligns with the fact of this document. Our context-augmented layout could capture such document-specific concept co-locations and adapt the initial text embedding in concept view to reflect the document context. To efficiently support the user to evaluate the summary quality from the perspective of concept appearance, we move the *concept glyph* with detailed document-summary information for each concept to a tooltip which can be triggered by hovering in iteration 2. This provides an effective overview and detail-on-demand exploration of the concepts in a document using interactive visual analysis.

5.2 Summary Evaluation

To facilitate the users to get an intuition about concepts from the document that are included in the summary compared the the concepts excluded from the summary (**R1**), we designed the *concept glyphs* (Fig. 1-2e) as described in Sec.5.1. Users can quickly filter out all but the "important" concepts, and then compare their distribution and context in the document and in the summary using the *concept glyphs* and the linked view to the document on the left (**R3**). To cater to user-specific analysis requirements (**R4**), we allow users to (1) choose the criteria (frequency or tf-idf) by which concepts should be considered "important" (Figure 1-2a), (2) choose the dimensionality reduction method (PCA, tSNE, or UMAP) to project the concepts (Figure 1-2b and 3), and filter them to only show the top K percent of concepts based on ConceptEVA's importance metric (Figure 1-2c).

Inspired by the experts' attempt to locate concepts with the "focus-on" function and their significant interest in it, we enhanced the "focus-on" function in iteration 2 to allow the user to switch perspectives and evaluate how well the current version of the summary addresses their specific areas of interest (Figure 5-c). When the user triggers the "focus-on" function, they will be able to select from full list of the concepts sorted by their appearance frequency in the original document (Figure 1-2d). Users can select one or



(a) Concept View in Iteration 1

(b) Default Layout of Concept View in Iteration 2

(c) "Focus-on" Layout of Concept View in Iteration 2

Figure 5: A comparison of the embedding-based layout in iteration 1 and the context-augmented latyout in iteration 2 for the concept view. All three figures show 80% of the concepts from the paper [40]. The circle size represents frequency (The size scale and ontology query parameters are slightly different between iteration 1 and 2). The "focus-on" layout in (c) focuses on the concept "algorithm".

multiple concepts based on their research interests and trigger a corresponding update of the concept view layout. The concept they choose to focus on will "float to the top" of the concept view, i.e., move to the top of the view, and the rest of the concept will "sink" to the bottom, with semantically or contextual-wise more relevant concepts pulled higher towards the top and less relevant concepts pushed lower towards the bottom. Meanwhile, the horizontal layout remains to reflect the concepts semantic and contextual distance determined by the user-chosen projection method. For instance, the layout in Fig. 5-c was focused on the concept "algorithm". We can see the related concepts including "methodologies", "validation", and "systems" are also pulled upwards. Meanwhile, the layout of the remaining concepts Fig. 5-b is locally maintained, continuing to reflect their semantic and contextual closeness in the document. This will further facilitate the concept selection and inform the customization task described in Sec. 5.3.

5.3 Summary Customization

Reflecting on the requirements we collect for a "good" summary (Sec 3), we approach summary customization in two ways: at a concept level, we see summary customization as determining what concepts are included when generating the summary, while at a structural level, we see it as inserting, reordering, and rewriting content. Users can achieve the concept-level summarization by selecting a group of concepts from the concept view to prioritize for the next version of the summary. Based on user selection, the summarizer extracts relevant sentences from the document as described in Sec. 4.2 and inputs them to the summarization pipeline for a customized summary that better addresses the concepts of interest.

The AI-generated summarization approach focuses more on the content than the flow of the summary, and was seen in the expert review as compromising the logical and narrative connection from one sentence to the next (see Sec. 6 for details). To address these concerns about the summary quality, we extended the interactions supported in ConceptEVA with an interactive summary editor to facilitate better human-AI collaboration in iteration 2. With the AI-generated summary as a starting point, the summary editor (Figure 1-3) helps users iteratively customize or extend the summary (**R1** & **R2**) by: (1) choosing from a list of candidate sentences for all user-selected concepts categorized by concept name, and inserting them into the summary, (2) updating a particular sentence in the summary with automatically paraphrased sentences generated with the paraphrasing model in Sec. 4.2 (Figure 1-3a), and (3) interactively editing, reordering, or deleting any sentences. In this way, a human-in-the-loop summary will be generated as the final output of the summary customization process in which user knowledge and judgments are effectively cooperated with the NLP techniques described in Section 4.2.

6 EXPERT REVIEW OF ITERATION 1

Iteration 1 of ConceptEVA was evaluated through expert review with three participants (2 male, 1 female). Given our prototype was backed with a NLP model more suited for scientific document analysis, we invited three experts with Ph.D. degrees in computer science with InfoVis as their research focus. Participant details are listed below, with years of experience in reading/reviewing academic papers included in parentheses.

- E1: software engineer (5–10 years).
- E2: senior applied scientist and former academic (10–20 years).
- E3: data scientist (5 to 10 years).

The review was conducted online via a video conference setting. Participants were first introduced to ConceptEVA's functions and features and given trial tasks with a test dataset to familiarize them with the interface.

Participants then used ConceptEVA to finish two open-ended tasks while following a concurrent think-aloud protocol: (1) verify the auto-generated summary for a given document, and (2) generate a customized summary according to a set of requirements provided to them. Since the participants were experienced researchers in infovis, we also collected their feedback and recommendations on the system as suggestions to incorporate into iteration 2. Iteration 1 was received positively in general, especially idea of evaluating a

document summary by examining the concepts (E1, E2, E3), context and support views to compare the document and the summary (E2, E3), but the quality of the generated summary was not considered sufficient (E1, E2, E3). Specific feedback is listed as follows:

- ***Concept Extraction & Separation:*** Concept identification through fuzzy matching between document terms and the reference ontology sometimes produced results that the experts (E1, E2, E3) found confusing. Iteration 1's implementation of the "focus-on" interaction was also not deemed helpful likely due to the issues concerning the fuzzy matching (E1, E2, E3), though all experts expressed considerable interest and pointed out potential ways for improvement. E1 and E2 also expressed that they expected a better-functioned "focus-on" tool with more intuitive interaction. E3 also suggested providing concept searching functions, showing the frequency of the concepts, and sorting the searching list accordingly.
- ***Information support:*** The visual representation of the concepts and the way they supported the comparison of the summary against the document was deemed helpful (E2, E3). Showing co-occurrence information of concepts was recommended (E1, E2, E3).
- ***Summary quality and presentation:*** An initial paragraph-like summary shown to E1 & E2 was deemed to not have a logical flow, while a bullet-point format change with E3 was received well. However, E3 was uncertain on how well they could "trust" the summary if it were of an unfamiliar paper, and recommended showing additional information to increase the user's confidence in the summary.

7 USER STUDY OF ITERATION 2

Lessons learned from the expert review helped focus the redesign of ConceptEVA and focus its evaluation through tasks that reflect how a researcher may approach summarizing an academic paper. Specifically, we decided to focus our study on whether and how a participant is able to generate a summary of a paper with which they are familiar using ConceptEVA such that the summary is relevant to their research interests.

While comparing the use of ConceptEVA with an existing summarization tool would be ideal, to our knowledge there is no existing summarization tool designed for research documents. We thus chose human-generated summaries by each participant as the baseline for that participant. While this means there is no "standard" baseline across all participants, this approach gives us better ecological validity as each participant would generate a summary that is relevant to their own interests and research contexts. Therefore, the current baseline for researchers would be to generate a summary by themselves—unaided by other tools. This would serve two purposes. Firstly, by generating their own summary manually, they gain familiarity with the document and are able to use ConceptEVA as a tool to refresh their memory, navigate the concepts relevant to the document, and be able to compare the summary they generate using ConceptEVA against their own manually-generated summary. Secondly, the process serves to emphasize our idea that ConceptEVA is *not* intended as a replacement for reading the document; it is intended to augment the way the document is explored.

This necessitated a study with a within-subjects component where each participant first generated a summary manually before attempting the same task on ConceptEVA. For the same reason, there was no counterbalancing: asking all participants to perform the manual summarization task first allowed us to ensure they were familiar with the document before they used ConceptEVA. It also allowed participants to critically examine the extent to which they could create a summary that was relevant to their own interest in the document. We used two test papers [43, 64], one for six participants in this study.

7.1 Participants

We recruited 12 participants (4 female, 8 male, aged 25–44 years), comprising 10 Ph.D. students, 1 university faculty, and 1 research engineer from a technology company. Seven participants reported they had been actively reading academic papers for 5-10 years, and the remaining five reported less than 5 years. And 10 participants reported they had written a summary/abstract/short description for an academic paper more than 10 times before the study, and the remaining two did it for 3-10 times. Two of the 12 participants reported themselves as native English speakers.

7.2 Experimental Setup

We conducted the study remotely considering the varied geographical locations of the participants and a safety measures surrounding the uncertain conditions of COVID-19. Instructions for the offline study task **T1** were shared with participants no less than 12 hours before the online study session began. For the online study session, the participants were asked to access ConceptEVAfrom a remote server and participate in the study with their own machine and external devices. Six participants used the Chrome browser with the Windows operating system, four used Chrome with MacOS, and the remaining two used the Safari browser with MacOS for the tasks. The setup, tasks, and durations were decided based on a pilot study with three participants: one native and two non-native English speakers.

We asked the participants to follow the "think aloud" protocol and audio- and video-recorded them during the task. Each participant received a $10 Amazon gift card as a compensation for their participation.

7.3 Summarization Guidelines

Based on findings from our survey of research practitioners explained in Sec. 3, we constructed a set of guidelines for participants to follow when generating a summary manually or using ConceptEVA. The guidelines were presented in the form of the following list of questions that participants could try and answer in their summary.

G1 ***Content.*** What is the paper about? What are the contributions?
G2 ***Approach.*** If the paper addresses a problem, how does it do it?
G3 ***Comparison.*** If the paper addresses a problem, how does its approach compare to existing approaches to address the same problem?
G4 ***Insights.*** What insights does the paper offer from its analysis or evaluation of the approach?

G5 ***Critique.*** What are the strengths and weaknesses of the approach?

G6 ***Implications.*** What are the implications of the work to your own interests and/or research?

We made it clear to participants that they were free to choose some, all, or even none of the guidelines below when generating the summary. In the procedure below, we would ask the participants which of the guidelines they followed for each summarization process: manual and using ConceptEVA.

7.4 Procedure

Each participant was provided with a research paper a few days in advance of the scheduled session with the study moderator, along with the guidelines listed in Sec.7.3. Each participant was then assigned the following tasks:

T1: Manual summarization.

- We asked participants to read the paper and manually generate a summary between a minimum of 100 and a maximum of 150 words reflecting what they found interesting in the paper. This summary was to be sent to the moderator in advance of their scheduled session. This represents the baseline for each participant, indicating the summary they would generate without ConceptEVA. It also ensures that participants read the paper before the start of the study.
- After their summary was received, participants were also asked to fill in a survey relating to their background and demographics. They were also asked to respond on a 7-point Likert scale (one for each guideline in Sec. 7.3) the extent to which they followed the guideline.
- Participants were also asked to report on their experience of the summarization task on the NASA TLX scale [20].

T2: Automated summarization.

- Participants were shown the automated summary generated without human intervention and asked to read through it.

T3: Human-in-the-loop summarization.

- Participants were introduced to the ConceptEVA interface and allowed to explore it through mini-tasks that reflected the process they would follow in their main task. This training/exploration session used a paper different from the one used for their tasks.
- Participants were then instructed to generate a summary of the same paper as in T1, following the same prompts and guidelines, but this time using ConceptEVA to explore and focus on concepts of interest and choosing relevant concepts to steer the summary generated. Throughout this exploration participants were instructed to follow a concurrent think-aloud protocol where they verbalized their thinking during their exploration.
- At the end of this process, they responded to a 7-point Likert scale (same as in T1) showing the extent to which they followed each guideline from Sec. 7.3.
- Participants reported on their experience of the summarization task on the NASA TLX scale.

T4: Rating all summaries.

- Participants finally rated on a 7-point Likert scale their satisfaction with (a) their manually-generated summary from T1, (b) ConceptEVA's automated summary with no human intervention from T2, and (c) the summary they generated in T3 using ConceptEVA by focusing on concepts of interest. They were allowed to re-read all three summaries before reporting on their satisfaction. The reason behind choosing "satisfaction" as a metric and for having participants rating their own summaries as opposed to others' summaries are related. Recall that the reason behind proposing ConceptEVA was that different readers of the same research article may emphasize different aspects when generating a summary of the paper. A participant with their own concepts of interest in a given paper would have takeaways that are influenced by these interests, which would in turn be reflected in their summary of the paper. We deemed that it would be less insightful for them to evaluate a summary generated by a different participant with different interests and takeaways. Instead, having the participant examine the summaries they have themselves created through three approaches could potentially reveal more insights into how well the human-in-the-loop approach has worked, as each participant can examine all summaries through the lens of their interest in the paper. For the same reason, "satisfaction" as a measure along with participant responses explaining the reasoning behind the rating allows us a way to understand what aspects of human-in-the-loop summarization are valuable for participants, albeit at the expense of specific insights more objective measures may provide.

The study did not focus on speed or quality of task performance, but on participants' own satisfaction with their experience and outcome. Thus task times were not restricted, and we did not track the time participants spent on Task 1, only their self-reported experience in writing the summary as described above. Participants in general spent between 60 and 90 minutes on tasks T2–T4.

8 RESULTS AND DISCUSSION

8.1 Summary Satisfaction

As mentioned in Sec. 7, we used each participant's manually-generated summary (**T1**) as a unique baseline for that participant. Ten of the 12 participants rated the automated summary (task **T2**) *lower* than the baseline, 8 out of 12 participants rated the summary generated using ConceptEVA's human-in-the-loop approach (task **T3**) *higher* than the baseline (Fig. 6). Recall that two papers were used in the study—6 participants summarized ConceptScope [64] and 6 summarized BodyVis [43]. Fig. 6 also includes a pointplot showing average ratings split across both papers. While the small participant pool makes it difficult to state with sufficient confidence whether participant satisfaction with the human-in-the-loop summarization using ConceptEVA is equivalent to their satisfaction with their own manually-generated summary, Fig. 6 suggests such an equivalence. In addition, a chi-squared test of independence showed a significant association between summarization approach and summary satisfaction rating, $\chi^2\,(8) = 23.5, p < 0.01$. On the other hand, a chi-squared test of independence showed no significant association between the paper used and summary satisfaction rating, $\chi^2\,(4) = 0.87, p = 0.93$. This indicates that the differences seen in Fig. 6 are more likely to be due to the summarization approach rather than the paper used in the task.

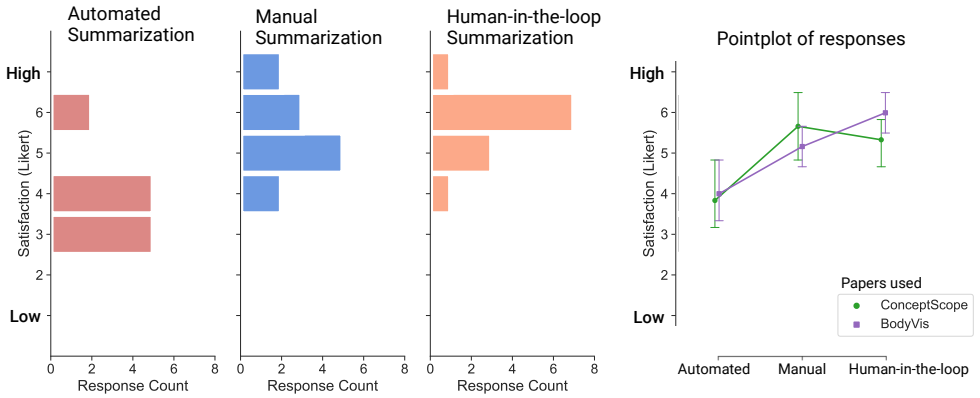


Figure 6: Distribution of participant responses on a 7-point Likert scale showing their level of satisfaction with the summaries from the automated approach in task T2, manual approach from task T1, and the human-in-the-loop approach in task T3 created with ConceptEVA. The three charts on the left show the distribution as response counts for each summarization approach. The right chart shows average values for each approach for the two papers used in the study, ConceptScope [64] and BodyVis [43], with error bars indicating 95% confidence intervals.

Participants who gave a higher rating for the human-in-the-loop approach reported being able to locate and focus on concepts more efficiently (P4, P6, P9), and on the content of the summary itself (P7). P7 observed that "*the contribution of this paper, was also well described in the* (human-in-the-loop generated) *summary.*" Participants who preferred the manual version of their summary to the human-in-the loop approach (P1, P3, P11) explained that they had their own idea of a summary that they wanted the generated version to reflect. For instance, P11 wanted the summary to focus on the paper methodology, and deleted all sentences from the automated summary, directing the system to pull new sentences from the paper focusing on "visualization", "concept", and "ontology". They proceeded to edit these new sentences based on their recall of the document and even manually wrote some text from scratch. These participants also reported a lower level of trust in the AI component of ConceptEVA through the study.

Participants' level of trust in the generated summary also appeared to be influenced by their confidence in their knowledge of the domains addressed in the paper. For instance, BodyVis [43], one of the papers used in the study, covers domains like participatory design, physiological sensing, and tangible learning, which the participants were relatively unfamiliar with. Their response to the summary generated by ConceptEVA was more positive. P4 reflected that *"in terms of ... describing the (BodyVis) system, maybe the one generated by ConceptEVA is kind of better... In the manually generated summary, although I put my focus there, I didn't do a good job like mentioning it. I don't think if I mentioned it."* P10 noted the automated summary addressed some of their own omissions: *"In my manual summary. I actually skipped some details, like I didn't really mention ... the feedback from children and the teachers (about BodyVis)."* In contrast, for a topic they were knowledgeable in, participants seemed to prefer their own interpretations and emphases, as P1 states: *"For the papers, if I already know that area, I have a certain expectation of what I need to look at. Then I would still prefer to write the summary by myself."*

In terms of the process, all participants reported being able to follow guidelines G1 (content) and G2 (approach) i.e., they rated themselves above 4 on a 7-point Likert scale. Six out of 12 participants reported being able to follow G4 (insights) and G6 (implications) as shown in Fig. 7. Participant ratings on being able to follow G3 (comparison) and G5 (critique) were skewed heavily toward the lower end of the scale. Participants P4 and P8 found it the most difficult to address these two guidelines, and they had a common approach: they attempted to find concepts related to "limitations" or "cons" to see the weaknesses reported in the paper itself and found this approach difficult to critique the paper and compare it with existing work. A low chance of success is expected with this approach as it is difficult to critique a paper by only examining the paper without a general sense of the related work. A summary that features such critique is difficult to automate as it would need knowledge as well as critical thinking about related work.

8.2 Summarization Experience

When responding to the NASA TLX scale (see Fig. 7) and rating their summarization experience, participants described the experience of using ConceptEVA as "*helpful*"(P1, P3, P7, P9, P11), "*useful*"(P1, P4, P6, P8, P9), "*amusing*" (P5) and "*enjoyable*" (P5). Eight participants reported that the concept view provided useful information such as the importance, appearance frequency, and co-occurrences of concepts. P4 and P6 also reported finding the focus-on function helpful to explore relationships with less dominant concepts. "*sometimes a concept is kind of minor...sheltered by those big circles...but by lifting it up you can see all the relation to other concepts. you can also like, and identify it directly.*" (P4).

The glyphs from the earlier iteration that were redesigned to be revealed only on detailed inspection were also deemed helpful by 6 participants, indicating perhaps that the glyph in isolation was helpful but several together were distracting. Since ConceptEVAwas

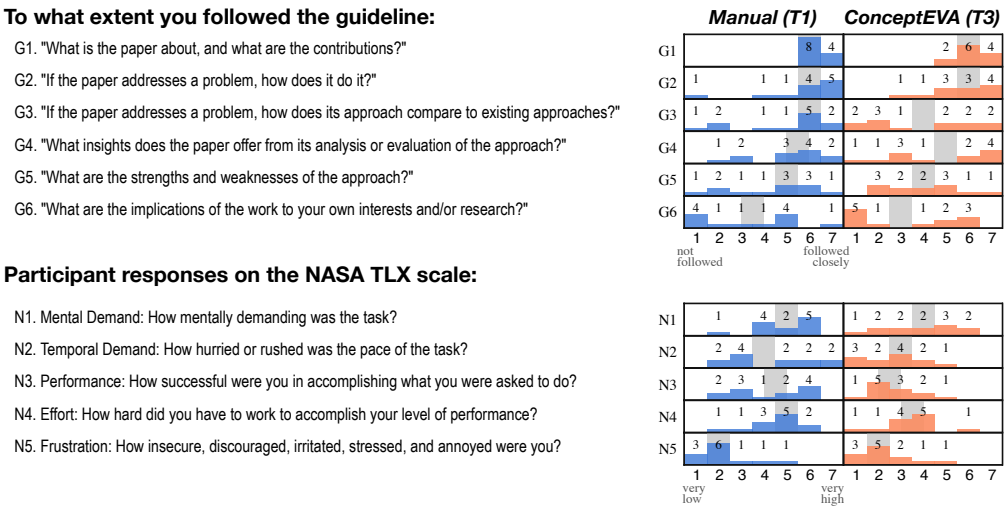


Figure 7: Ratings for manually-generated summary in T1 and human-in-the-loop summary in T3. Median ratings are in gray.

implemented for the browser, we also observed participants incorporating built-in browser functionalities such as search, translation (for bilingual/multilingual participants), and grammar checkers.

Participants also expressed their frustration when they "*can't find anything useful about the word they identified*"(P3) or "*lose the full picture of the paper*"(P8). Identifying relevant concepts is a function of the ontology, and a balance between the specificity of domain ontologies and the breadth of a general ontology such as DBpedia. On the other hand, issues related to identifying strengths and weaknesses of the work that may not be explicitly stated in the paper—echoing participant experiences described in Sec. 8.1—may be possible to address by additional visualization of document affect and sentiment [27].

8.3 Influence Factors on User Experience

When conducting tasks **T3** and **T4**, we observed four dominant factors that appeared to influence participants' use and preferences of certain functionalities in ConceptEVA. Please refer to Table 3 in the supplementary materials for a detailed report of participants' behavior patterns when performing the user study tasks.

- *Academic reading experience and skill influences exploration.* Participants such as senior PhD students and faculty/researchers preferred to read the original text of the paper. P5, a graduate student with 5–10 years of experience reading academic papers, said they preferred to read the original text of the paper, but also said that the concept view *"is actually really good with the way my brain is... I just think of words, and then it (the focus-on function) has the words I want. This kind of maps with my thinking, which is very amusing."* In contrast, participants with either less academic experience or from a different domain found direct text reading difficult. For example, P7 thought the paper reading process was"*very overwhelming*" while P8 reported that they "*don't have the full picture of the paper in this way*". They preferred to use the visualizations—the projection view or the Focus-on too—to get a high-level overview, and then "*grab information based on the concept that I'm giving*"(P11). While this is part of the intention behind designing the visualizations (esp. **R3**), a longitudinal study may be needed to explore how ConceptEVA may be used as a way to scaffold students' ability to read and understand academic text. Note that P5, P7, P8, and P11 are all graduate students, but P5 identifies as a native English speaker while the others do not. While this may not be the reason for the difference, it brings up the issue of reading skill, a factor that was not evaluated in the study.
- *Academic writing experience influences summarization.* An extension of the above observation means that participants' academic writing experience would influence how they used ConceptEVA to summarize text. P5 found the workflow afforded by ConceptEVA useful, and that it was "*doing most of the work for me*", such as "*constructing sentences I would put in my paper, or something letting me take what either my problem is or what I'm thinking about looking at the paper, and like merging these things together*". They also appreciated "*the freedom of allowing more editing*" in the summary editing panel (**R4**), and used it to directly edit the summary sentences. P10 reported finding it useful to "*pull out the related sentences categorized by each of the concepts you selected*" (**R3**). Other experienced participants like P11 reported that ConceptEVA "*doesn't encode (sic) their standard of generating the summary.*" Note that P11 is also the participant who heavily edited the generated summary (Sec. 8.1).
- *Domain familiarity influences use of ConceptEVA.* Participants' reflections indicated that their knowledge of the domain covered in the document would influence how they would use ConceptEVA. P1 mentioned that "*if I'm reading a machine learning paper or deep learning one that I'm not quite familiar with (the domain)*", they would prefer to use the concept view to "*understand what kind of concepts they (the paper) have*" and would like to see definitions of the concept in ConceptEVA. On the other hand, for documents in their own domain, they said they would *"have a certain expectation of what I need to look at. Then I would still prefer to write the summary by myself.*" This was also seen in P8's

approach in the study: they were unfamiliar with the paper they were asked to read and requested more information support as they did not have "*a general picture of the paper.*"

- *Mental map of document influences use of visual interface.* Eleven of the 12 participants reported being happy with the visual interface for the summary customization task. While distributing their time to the three panels in ConceptEVAin different ways, 11 out of the 12 participants embraced the visual interface for the summary customization task in our study. Participants preferred different aspects of the interface depending on the way they approached ideas in the paper. P5, quoted earlier in this section, stated how the concept view layout mirrored the way they think. P11, on the other hand, preferred the "paper info" panel to the concept view *"because I can see I know where it (the concept) is (in the paper)."* They even chose to search the concept directly in the PDF version of the paper after briefly exploring the Focus-on function in the concept view panel, explaining that "*it's quite a huge number of information ... it's a little bit hard to draw the connection between the information inside the original paper and the (concept view) exploration panel. That's why I just ignore the exploration panel.*" Others found the paper info panel disorienting as it provided a view of the paper that was different from the PDF layout they had initially read, stating, *"I don't have, like the mental map of the original pdf. It's gone"* (P5), *"Here everything's like um very flat. So I don't know where it is."* (P12), and ' *'I didn't use this. Yeah, this part was well overwhelming"* (P7).

8.4 Limitations and Future Work

One of the issues that came up through the iterations is striking the right balance between the use case scenarios of ConceptEVA, specifically its use to explore a paper as an alternative to reading. Similar "distant reading" approaches in the social sciences have received criticism for being suggested as objective alternatives to close reading, a practice considered integral to scholarship [1]. In our studies, the expert review evaluation for the first iteration of ConceptEVA did not require participants to read the paper in advance. Thus they spent more time using the system to understand the paper content—which was not the main focus of the system—than to generate and evaluate the summary. The study setup following iteration 2 ensured that participants were already familiar with the paper, which allowed them to focus on the summary evaluation and customization tasks. Participant reflections we saw in Sec. 8.1 and Sec. 8.3 show that participants still used ConceptEVA as a way to check if they missed any important concepts, especially if they were unfamiliar with the domain of the paper. Participant P3 suggested using ConceptEVA as a way to skim through papers so that "*if frequent concepts are not what I care, I can just leave this paper and turn to others.*" On the other hand, comments about the disorienting effect of the paper layout in the paper info panel (see "mental maps" in Sec. 8.3) indicates that a better application of ConceptEVA would be toward supporting and summarization and *verification*, rather than exploration. Integral to this approach would be to design a paper information view that preserves the appearance of the PDF view, thus preserving the reader's mental map and allowing them to build upon their close reading of the paper.

The two test papers we chose for the user study were corresponding to the two different conditions—highly interdisciplinary papers spanning at least five domains and relatively typical CHI papers describing the algorithm, user study, and visualization design. Because of the authors' limited knowledge background, we chose two CHI papers in which we had a better understanding and control of the content for our user study. We will eliminate this limitation by testing ConceptEVA on more diverse papers in the future. Besides, we are aware of the different summarization complexity for papers from different domains [23, 48, 60, 61], but consider it more of an NLP research problem rather than our main focus.

Participants also made suggestions for additional functions and features. The most popular suggestions fell under the category of richer view coordination between the panels. Specifically, participants suggested being able to support concept provenance and filtering within a selected section, or a direct linking between the summary text and the paper information panel. However, this would also mean that ConceptEVA becomes more of an exploration tool providing an alternative to reading the paper rather than a support to summarize a paper, which is a different scope of work altogether, and a requirement that needs closer examination in terms of benefits and pitfalls. On the other hand, other suggestions such as the one by P1 about being able to group concepts into groups relevant to the summary such as "definition", "pipeline", and "preprocessing method". While the groups listed by P1 might work for a data science or data visualization domain, other domains might require entirely different groups than can then be examined to summarize contributions, offer critique, and present other salient ideas. Allowing the user to create custom groups aided by additional NLP approaches like sentiment analysis and topic modeling could help users reflect on and critique the paper, and can be a helpful function to consider in a future iteration of the work.

Finally, a limitation of our study include technical issues such as network delays, rendering performance issues, and back-end computations to update concept embedding, sentence paraphrasing, or summary generation itself. These, when they occurred, resulted in latency that influenced participants' experience and potentially their responses to questions like the NASA TLX scale. While the focus of this work is not engineering or optimisation of the system, our future iterations will attempt to cut down performance or networking issues relating to latency.

9 CONCLUSION

We have presented ConceptEVA, an interactive document summarization system aimed at long, and multi-domain documents of the kind seen in academic publications. We show the iterative development and evaluation of ConceptEVA through two iterations. The first iteration incorporates a hierarchical summarization technique with an interactive visualization of concepts extracted from the document using a reference ontology. The second iteration, developed after evaluating the first iteration through an expert review, incorporates a multi-task longformer encoder decoder pre-trained for scientific documents that we fine-tune for paraphrasing and sentence embedding to handle long documents, and concepts visualized using a force-directed network that preserves semantic as well

as co-occurrence relationships of document concepts. We also introduce a "focus-on" function that allows users to choose concepts of interest, examine their relationship with co-occurring concepts, and choose relevant concepts that will then be incorporated into a custom summary. An evaluation of ConceptEVA's second iteration through a within-subjects study using manually-generated summaries as baseline shows that ConceptEVA was helpful to participants for content-specific aspects of summarization, but participants with less experience struggled with critique-related aspects of summarization. Participants largely preferred the summary created through ConceptEVA's human-in-the-loop approach over their own manually-generated summaries. We discuss the implications of our findings and suggest future development and evaluations of mixed-initiative summarization systems.

ACKNOWLEDGMENTS

We thank the participants of our studies and the anonymous reviewers for their feedback and suggestions. This research is sponsored in part by Bosch Research and the National Science Foundation through grant ITE-2134901.

REFERENCES

[1] Maurizio Ascari. 2014. The Dangers of Distant Reading: Reassessing Moretti's Approach to Literary Genres. *Genre: Forms of Discourse and Culture* 47, 1 (2014), 1–19. https://doi.org/10.1215/00166928-2392348

[2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*. Springer, Berlin, Heidelberg, 722–735.

[3] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *CoRR* abs/2004.05150 (2020), 17. https://arxiv.org/abs/2004.05150

[4] Matthew Berger, Katherine McDonough, and Lee M Seversky. 2016. cite2vec: Citation-driven document exploration via word embeddings. *IEEE transactions on visualization and computer graphics* 23, 1 (2016), 691–700. https://doi.org/10.1109/TVCG.2016.2598667

[5] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD international conference on Management of data*. ACM, New York, NY, 1247–1250. https://doi.org/10.1145/1376616.1376746

[6] Rishi Bommasani and Claire Cardie. 2020. Intrinsic Evaluation of Summarization Datasets. In *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online, 8075–8096. https://doi.org/10.18653/v1/2020.emnlp-main.649

[7] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022. Improving Language Models by Retrieving from Trillions of Tokens. In *Proceedings of the International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, Baltimore, Maryland, 2206–2240.

[8] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 632–642.

[9] Senthil Chandrasegaran, Chris Bryan, Hidekazu Shidara, Tun-Yeng Chuan, and Kwan-Liu Ma. 2019. TalkTraces: Real-Time Capture and Visualization of Verbal Content in Meetings. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 577:1–577:14. https://doi.org/10.1145/3290605.3300807

[10] Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. 2019. On the Use of ArXiv as a Dataset. *CoRR* abs/1905.00075 (2019), 7. https://arxiv.org/abs/1905.00075

[11] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2270–2282. https://doi.org/10.18653/v1/2020.acl-main.207

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. , 4171–4186 pages. https://doi.org/10.18653/v1/N19-1423

[13] Bonnie Dorr, Christof Monz, Stacy President, Richard Schwartz, and David Zajic. 2005. A methodology for extrinsic evaluation of text summarization: does ROUGE correlate?. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Association for Computational Linguistics, Ann Arbor, Michigan, 1–8.

[14] Mennatallah El-Assady, Valentin Gold, Carmela Acevedo, Christopher Collins, and Daniel Keim. 2016. ConToVi: Multi-party conversation exploration using topic-space views. *Computer Graphics Forum* 35, 3 (2016), 431–440. https://doi.org/10.1111/cgf.12919

[15] Mennatallah El-Assady, Rita Sevastjanova, Bela Gipp, Daniel Keim, and Christopher Collins. 2017. NEREx: Named-Entity Relationship Exploration in Multi-Party Conversations. *Computer Graphics Forum* 36, 3 (2017), 213–225. https://doi.org/10.1111/cgf.13181

[16] Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications* 165 (2021), 113679. https://doi.org/10.1016/j.eswa.2020.113679

[17] Som Gupta and Sanjai Kumar Gupta. 2019. Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications* 121 (2019), 49–65. https://doi.org/10.1016/j.eswa.2018.12.011

[18] Vishal Gupta and Gurpreet Singh Lehal. 2010. A survey of text summarization extractive techniques. *Journal of emerging technologies in web intelligence* 2, 3 (2010), 258–268. https://doi.org/10.4304/jetwi.2.3.258-268

[19] Shanmugasundaram Hariharan and Rengaramanujam Srinivasan. 2010. Studies on intrinsic summary evaluation. *International Journal of Artificial Intelligence and Soft Computing* 2, 1-2 (2010), 58–76. https://doi.org/10.1504/IJAISC.2010.032513

[20] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Human Mental Workload*. Advances in psychology, Vol. 52. Elsevier, Amsterdam, Netherlands, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[21] Florian Heimerl and Michael Gleicher. 2018. Interactive analysis of word vector embeddings. *Computer Graphics Forum* 37, 3 (2018), 253–265. https://doi.org/10.1111/cgf.13417

[22] Tsutomu Hirao, Yutaka Sasaki, and Hideki Isozaki. 2001. An extrinsic evaluation for question-biased text summarization on QA tasks. In *Proceedings of the NAACL Workshop on Automatic Summarization*. Association for Computational Linguistics, Pittsburgh, PA, 61–68.

[23] Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. 2021. Summarization of legal documents: Where are we now and the way forward. *Computer Science Review* 40 (2021), 100388. https://doi.org/10.1016/j.cosrev.2021.100388

[24] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572

[25] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through Memorization: Nearest Neighbor Language Models. In *International Conference on Learning Representations*. OpenReview.net, Addis Ababa, Ethiopia, 13.

[26] Dae Hyun Kim, Enamul Hoque, Juho Kim, and Maneesh Agrawala. 2018. Facilitating document reading by linking text and tables. In *Proceedings of the ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, 423–434. https://doi.org/10.1145/3242587.3242617

[27] Kostiantyn Kucher, Carita Paradis, and Andreas Kerren. 2018. The state of the art in sentiment visualization. *Computer Graphics Forum* 37, 1 (2018), 71–96. https://doi.org/10.1111/cgf.13217

[28] Vivian Lai, Alison Smith-Renner, Ke Zhang, Ruijia Cheng, Wenjuan Zhang, Joel Tetreault, and Alejandro Jaimes-Larrarte. 2022. An Exploration of Post-Editing Effectiveness in Text Summarization. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 475–493. https://doi.org/10.18653/v1/2022.naacl-main.35

[29] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

[30] Haoran Li, Junnan Zhu, Jiajun Zhang, Chengqing Zong, and Xiaodong He. 2020. Keywords-Guided Abstractive Sentence Summarization. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 05 (2020), 8196–8203. https://doi.org/10.1609/aaai.v34i05.6333

[31] Allen Yilun Lin, Joshua Ford, Eytan Adar, and Brent Hecht. 2018. VizByWiki: Mining data visualizations from the web to enrich news articles. In *Proceedings of the World Wide Web Conference*. ACM, New York, NY, 873–882. https://doi.org/10.1145/3178876.3186135

[32] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81.

[33] Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*. Association for Computational Linguistics, Online, 495–501. https://doi.org/10.3115/990820.990892

[34] Shusen Liu, Peer-Timo Bremer, Jayaraman J Thiagarajan, Vivek Srikumar, Bei Wang, Yarden Livnat, and Valerio Pascucci. 2017. Visual exploration of semantic relationships in neural word embeddings. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 553–562. https://doi.org/10.1109/TVCG.2017.2745141

[35] Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development* 2, 2 (1958), 159–165. https://doi.org/10.1147/rd.22.0159

[36] Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *CoRR* 1802.03426 (2018), 63. https://doi.org/10.48550/ARXIV.1802.03426

[37] Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia Spotlight: Shedding Light on the Web of Documents. In *Proceedings of the ACM International Conference on Semantic Systems*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/2063518.2063519

[38] Derek Miller. 2019. BERT Extractive Summarizer. https://github.com/dmmiller612/bert-extractive-summarizer.

[39] Francesco Moramarco, Alex Papadopoulos Korfiatis, Aleksandar Savkov, and Ehud Reiter. 2021. A Preliminary Study on Evaluating Consultation Notes With Post-Editing. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*. Association for Computational Linguistics, Online, 62–68.

[40] Tamara Munzner. 2009. A nested model for visualization design and validation. *IEEE transactions on visualization and computer graphics* 15, 6 (2009), 921–928. https://doi.org/10.1109/TVCG.2009.111

[41] Gabriel Murray, Thomas Kleinbauer, Peter Poller, Tilman Becker, Steve Renals, and Jonathan Kilgour. 2009. Extrinsic summarization evaluation: A decision audit task. *ACM Transactions on Speech and Language Processing* 6, 2 (2009), 1–29. https://doi.org/10.1145/1596517.1596518

[42] Arpit Narechania, Alireza Karduni, Ryan Wesslen, and Emily Wall. 2021. vitaLITy: Promoting Serendipitous Discovery of Academic Literature with Transformers & Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 486–496. https://doi.org/10.1109/TVCG.2021.3114820

[43] Leyla Norooz, Matthew Louis Mauriello, Anita Jorgensen, Brenna McNally, and Jon E Froehlich. 2015. BodyVis: A new approach to body learning through wearable sensing and visualization. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1025–1034. https://doi.org/10.1145/2702123.2702299

[44] Haekyu Park, Nilaksh Das, Rahul Duggal, Austin P Wright, Omar Shaikh, Fred Hohman, and Duen Horng Polo Chau. 2021. NeuroCartography: Scalable Automatic Visual Summarization of Concepts in Deep Neural Networks. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 813–823. https://doi.org/10.1109/TVCG.2021.3114858

[45] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

[46] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551. https://doi.org/10.5555/3455716.3455856

[47] Erhard Rahm and Philip A Bernstein. 2001. A survey of approaches to automatic schema matching. *the VLDB Journal* 10, 4 (2001), 334–350. https://doi.org/10.1007/S007780100057

[48] Moiz Rauf, Sebastian Padó, and Michael Pradel. 2022. Meta Learning for Code Summarization. *CoRR* 2201.08310 (2022), 5. https://doi.org/10.48550/arxiv.2201.08310

[49] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, Hong Kong, China, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[50] Yves Scherrer. 2020. TaPaCo: A Corpus of Sentential Paraphrases for 73 Languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 6868–6873.

[51] Tian Shi, Yaser Keneshloo, Naren Ramakrishnan, and Chandan K Reddy. 2021. Neural abstractive text summarization with sequence-to-sequence models. *ACM Transactions on Data Science* 2, 1 (2021), 1–37. https://doi.org/10.1145/3419106

[52] Daniel Smilkov, Nikhil Thorat, Charles Nicholson, Emily Reif, Fernanda B Viégas, and Martin Wattenberg. 2016. Embedding projector: Interactive visualization and interpretation of embeddings. *CoRR* 1611.05469 (2016), 4. https://doi.org/10.48550/arxiv.1611.05469

[53] Ji Soo Yi, Rachel Melton, John Stasko, and Julie A Jacko. 2005. Dust & magnet: multivariate information visualization using a magnet metaphor. *Information visualization* 4, 4 (2005), 239–256. https://doi.org/10.1057/palgrave.ivs.9500099

[54] Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. Abstractive document summarization with a graph-based attentional neural model. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 1171–1181.

[55] Michael E. Tipping and Christopher M. Bishop. 1999. Mixtures of Probabilistic Principal Component Analyzers. *Neural Computation* 11, 2 (02 1999), 443–482. https://doi.org/10.1162/089976699300016728

[56] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605.

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017), 11.

[58] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Brussels, Belgium, 353–355. https://doi.org/10.18653/v1/W18-5446

[59] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 1112–1122.

[60] Jeff Wu, Long Ouyang, Daniel M Ziegler, Nissan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively Summarizing Books with Human Feedback. *CoRR* 2109.10862 (2021), 37. https://doi.org/10.48550/arXiv.2109.10862

[61] Min Yang, Qiang Qu, Ying Shen, Qiao Liu, Wei Zhao, and Jia Zhu. 2018. Aspect and Sentiment Aware Abstractive Review Summarization. In *Proceedings of the International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 1110–1120.

[62] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*. PMLR, online, 11328–11339.

[63] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the International Conference on Learning Representations*. Openreview.net, online, 43.

[64] Xiaoyu Zhang, Senthil Chandrasegaran, and Kwan-Liu Ma. 2021. ConceptScope: Organizing and visualizing knowledge in documents based on domain ontology. In *Proceedings of the ACM CHI conference on human factors in computing systems*. ACM, New York, NY, 19:1–19:13. https://doi.org/10.1145/3411764.3445396

[65] Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 1298–1308. https://doi.org/10.18653/v1/N19-1131