Handling Comments in Collaborative Documents through Interactions

Anonymous ACL submission

Abstract

Comments are widely used by users in collaborative documents every day. The documents' comments enable collaborative editing and review dynamics, transforming each document into a context-sensitive communication channel. Understanding the role of comments in communication dynamics within documents is the first step towards automating their management. In this paper we propose the first ever taxonomy for different types of in-document comments based on analysis of a large scale dataset of public documents from the web. We envision that the next generation of intelligent collaborative document experiences allow interactive creation and consumption of content, there We also introduce the components necessary for developing novel tools that automate the handling of comments through natural language interaction with the documents. We identify the commands that users would use to respond to various types of comments. We train machine learning algorithms to recognize the different types of comments and assess their feasibility. We conclude by discussing some of the implications for the design of automatic document management tools.

1 Introduction

Comments on collaborative documents serve as a communication channel. This type of context-specific communication allows dynamics to review and edit content within the document. Collaborative text editors have visual components that allow users to associate a comment with a specific part of the content. This provides additional context in situations where the conversation focuses on a specific part of the document (Churchill et al., 2000). As we can see, the amount of contextualization in communication that document comments permit is too complex and costly to recreate in other communications means outside of a document. For example, a request for changing a certain part of a document's content (e.g. a paragraph's sentence) through email would require much additional information to be provided about all of the context before requesting the change.

In this paper, we present a novel taxonomy of the types of comments detected in a collection of public documents. We detect three main categories of intents for comments that are Modification, Information Exchange, and Social Communication. We show that supervised models can successfully be trained to identify the type of comments. We conducted additional studies where users provided commands for resolve each type of comment. Users were asked to provide commands the way they would when interacting with a voice assistant through natural language. We find the most common commands as well as their structure. The following summarizes our contributions:

1. Using a large-scale public document dataset that we have curated and release with this paper, we analyze the role of document comments and propose a taxonomy of comments' intents and sub-intents.
2. We propose methods for determining the intent of comments and discuss their potential for automation.
3. We analyze how people would handle each type of comment by providing voice commands.

The paper continues with the following structure. In section two, we describe the previous work in this area of study. In section three, we describe the dataset collection. In section four, we explain the process of identifying the intents. In sections five and six, we present the results of the two case studies, followed by section seven, where we discuss them. We conclude the paper with the conclusions of the work in section eight.

Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering, pages 176 - 186

2 Related Work

2.1 Comments management on collaborative documents

Collaborative document editing has been present since the appearance of web 2.0, which implied a paradigm shift. Web 2.0 allowed that the task of adding content to the web was not an exclusive activity of the webmaster (Lewis, 2006). The dynamics of collective contribution and curation required the implementation of tools that coordinated the processes of preparation, evaluation, and production of the information. Wikipedia was one of the pioneering platforms in implementing collective content production tools. The implementation of a communication channel in Wikipedia allowed asynchronous communication between users with different roles. Yang et al. studied the different types of comments and the functions that comments enable on Wikipedia (Yang et al., 2017). In the context of email messages, Dabbish et al. identified the common intents in the workplace (Dabbish et al., 2005). In this work, we study the taxonomy of comments in collaborative documents.

2.2 Intent classification

Understanding the intentions of users is a required task in multiple Natural Language Processing (NLP) applications. An example is chatbots, which after interpreting the intents and entities, are capable of responding to an unstructured message. Previous work has studied how intent detection techniques based on neural networks models often overcome classical methods (Khattak et al., 2021). Some activities require more context; for example, identifying the intent of an email only by the subject could be imprecise if we do not take into account the body of the email. Wang et al. explored how to detect the intent of an email based on the title and body of the email (Wang et al., 2019). In the case of collaborative documents, there are multiple elements that contribute to the context, such as the selected text, the paragraph text, and the comment text. In this work, we study intent detection models that use multiple elements of the context.

2.3 Voice commands for document editing

The effective management of document comments requires a reliable interpretation of voice commands and a clear understanding of user intents.

Using voice as an input interface is not something novel. In 1976, Reddy reviewed the effectiveness of acoustic, phonetic, syntactic, and semantic subsystems (Reddy, 1976). Some pioneering work detecting commands from audio include techniques where sequences of phonemes (Halle and Stevens, 1962) and prosodemes (Peterson, 1961) were interpreted as commands. The human voice is especially challenging to detect because of the variability among individuals (Radha and Vimala, 2012). Early work in human voice processing was constrained to a limited set of words (Pieraccini and Director, 2012). The feature engineering techniques over audio help to identify descriptors that characterize words. Some toolkits that extract a variety of those features emerged, such as SMILE (Eyben et al., 2010). These enabled some approaches based on classic machine learning techniques such as Support Vector Machines (Kanth and Saraswathi, 2015). The major change in performance and efficiency happened when neural networks were fed large amounts of data. Some early neural network approaches used Hidden Markov Models to detect words in English (Aldarmaki et al., 2021). Latest work in this field uses Transformers for detecting multi-speaker speech recognition (Chang et al., 2020).

2.4 Assisted Document Management

Assistance over document writing is an antique practice. Scribes were people who made copies and wrote letters on behalf of others not only to avoid the need to write for themselves but also because of illiteracy (Anzelc et al., 2021). The rules that humans use to transcribe text are often implicit and subjective. The automation of this process requires a first standardization effort; this explains why some speech-to-text tools include a commands sheet. There is a trend that dictation tools recognize more and more natural language. The latest approaches in automatic transcription (Gupta et al.) have moved away from providing a list of commands and now try to infer based on context. Nowadays, editing tools are not only designed to share information but also promote collaboration. Exchanging comments in a document is a communication channel widely used in companies and at a personal level. Our work extends on previous work that has enabled mechanisms to understand commands from natural language applied to document comments management.

3 Document Comments Dataset

This section explains the details of our process for preparing the document comment dataset. We have curated a set of documents that contain multiple comments from public sources available on the web. To our knowledge, there are currently no datasets available that have been curated for investigation of in-document comments. It is evident that such dataset is needed for research. Although it is possible to investigate comments on public pages such as Wikipedia, Reddit, Twitter, YouTube, or other web forums, however, their use case of comments on these forums is inherently very different than in-document comments used for collaborative authoring. In-document comments are interactive and conversational and commonly request and result in changes and updates to the content of the document that is shared. In-document comments are intended to be carefully reviewed by the intended recipients, and authors and reviewers tend to resolve and remove them prior to releasing documents to the public readers. This practice makes it very difficult to come across in-document comments in public mature documents. Private files which are earlier in the editing life-cycle are more likely to have threads of comments. We use public documents because releasing private files is not possible due to copyright and privacy concerns. In addition to the challenges mentioned, we observed that only a certain percentage of word documents from recent years (after 2003) support comments and that we were able to extract comments from them.

3.1 Data Collection

We used an initial index of 1,000,000 word documents from the web through the CommonCrawl (Com, a) and filtered them based on the language to obtain English 'en' documents from the index. We also filtered this collection to include only Microsoft Word documents with the **'.docx'** extension. The reasons behind the decision to use only .docx file were that 1) the non-binary nature of the XML files contained in the .docx bundle make the data extraction easy with common XML tools; and 2) in 2003 (the same year that the .docx format was introduced) the comments were integrated to the document interface.

We observed that Some files were duplicates of one another even though they were indexed at different addresses and had different filenames and URLs. For some instances, this was because of the changes between CommonCrawl index batches. In order to be able to detect duplicates of files and prevent duplicates from reappearing in our dataset, we compared their MD5 hashes with one another. We then addressed the issue for files that were not duplicates of one another but rather incremental versions; in those cases, we kept the document with a higher number of comments.

Through applying these constraints, we ended up with 107,885 total indexed .docx files in English dating between June 2013 and July 2020. Only 1,313 documents out of the 107,885 total indexed .docx files had comments (1.2% rate) and the final dataset contains 12,253 comments extracted from this set of 1,313 documents.

3.2 Data Processing

We use scripts via XML parser to extract the Microsoft Word meta-information about the document and each comment. For each comment, we extracted the information of its anchored paragraph, text selection, comment content, and responses to the comments. We once again filtered the documents using the inferred language provided by Microsoft Office to ensure they were in English. We preferred not to have to translate to prevent change in context and meaning through automatic translations. We anonymized the users' names and removed any personal identifying information to comply with ethical guidelines.

The complete dataset can be downloaded from the project's GitHub repository [1].

4 Document Comment Intents

4.1 Identification of intents

We use grounded theory to detect the different types of intentions present in the comments of the documents. We identified the following 3 general categories: **Modification**, **Information Exchange**, **Social Communication**.

4.2 Document Comments Annotation

A set of 5000 randomly selected comments were annotated by three coders. The annotators were sourced through the company KarmaHub.The interface for annotating comments is depicted in Figure 1. The green text highlights a sentence in the comment to be annotated, while the yellow text highlights the selected text associated with the comment. Two annotators selected intents and sub-intents for

[1] available upon publication

[width=]sentence.pdf

Figure 1: Annotation interface that shows the sentence to annotate (green) and its associated text (yellow). Annotators chose intents and sub-intents in the annotation area (orange).

each message, and a third annotator served as a tiebreaker, selecting the most accurate labels in cases of disagreement. We obtain a significant Kappa score of 0.65 for the agreement between annotators. The distribution of comments across sub-intents in the dataset is shown in Table 1.

We enabled an "Other" category when they were unable to identify the intent (i.e., a multilingual comment) or when the comment contained an intent not defined in our list. Only 297 (5.9 percent) of comments were classified as "Other."

5 Case Study 1 - Document Comments Classification

In this case study, we use labeled comments to train machine learning models and evaluate their performance. The evaluation of the trained models helps to validate their feasibility to be implemented in real-world solutions.

5.1 Classification Methods

We implement classical methods of machine learning as well as deep learning for the training of models that can classify intents. For the evaluation of classical models, we use the Supported Vector Machine (SVM) and Logistic Regression (LR) models. Additionally, we implemented classification models based on the Transformers (Vaswani et al., 2017) architecture. The distilled versions of BERT (Sanh et al., 2019) RoBERTa (Liu et al., 2019), and BART (Lewis et al., 2019) were fine-tuned with our data.

Adding fragments of texts that give context to the comments could influence their performance. The text elements that we consider are the following:

- Comment: The whole comment.
- Sentence in a comment: A single sentence of a comment.
- Selected text: The text to which the comment refers.
- Paragraph text: The text of the paragraph where the comment belongs.
- Thread text: The comments that precede the comment to be evaluated.

5.2 Classification Results

The training of the models was carried out at different hierarchical levels of categories. For each model, the text of the comment was evaluated as well as texts located in other regions of the document that correspond to the context. Table 3 shows the top category level performance metrics over all the data across models. From the results, we can see that the Transformer models had a similar overall performance.

The models were trained with a combination of context elements. Table 4 shows that there were no major changes to how context items can improve the classification task for comments. Transformer-based models accomplished this task with similar results across all models.

Performance across categories may vary depending on the hierarchy level of each category. Table 5 shows the results of how the models perform in the two top levels. The results show that the categories of Modification and Information Exchange and their subcategories maintained a similar performance, while the categories of social communication obtained a lower performance.

6 Case Study 2 - Voice Commands

Interacting with documents via voice is not something novel. Voice has enabled for years hands-free interactions while consuming or editing documents. Its usage is not limited to performance or accessibility scenarios; the emergence of virtual voice assistants has enabled new multi-device and multi-modal interactions.

Using voice to express ideas is a natural interaction between humans, but it adds extra complexity to machines. Peripherical input devices as keyboards convert electrical impulses to single characters; it reduces errors to user motricity or device mechanical-related issues. Machines rely on speech recognition algorithms to get accurate input from the voice. Even today, with sophisticated algorithms and huge volumes of data, the results are far from perfect. Being able to develop voice-based solutions implies dealing with uncertain information—the variability of ways to express the same concept help applications to be resilient to unexpected inputs.

Document dictation is one of the tasks that speech recognition enables. Dictation implies transcribing what is said to the document. To get syntactically correct results, these tools

Table 1: Document comment taxonomy.

Main	Level 1	Level 2	Level 3	Level 4
MODIFICATION (1883, 37.7%)	REQUEST (1611, 85.5%)	CONTENT (1209, 75%) / FORMAT (402, 25%)	EXPLICIT (1519, 94.2%) / NOT EXPLICIT (92, 5.8%)	ADD (835, 51.9%) / CHANGE (583, 36.1%) / DELETE (193, 12%)
	EXECUTION STATUS (272, 14.5%)	DONE (254, 93.3%) / PROMISE (18, 6.7%)		
INFORMATION EXCHANGE (2477, 49.7%)	PROVIDED (1771, 71.5%)	CONTEXT (1420, 80.1%) / REFERENCE (351, 19.9%)	POTENTIAL CHANGE (1104, 62.3%) / NOT POTENTIAL CHANGE (667, 37.7%)	
	REQUESTED (706, 28.5%)	ASKING DETAILS (554, 78.4%) / REQUESTING CONFIRMATION (152, 21.6%)	POTENTIAL CHANGE (600, 84.9%) / NOT POTENTIAL CHANGE (106, 15.1%)	
SOCIAL COMMUNICATION (343, 6.8%)	ACKNOWLEDGMENT (25, 7.2%)			
	DISCUSSION (143, 41.6%) / FEEDBACK (175, 51.2%)	CONTENT (174, 50.7%) / THREAD (144, 49.3%)	POTENTIAL CHANGE (117, 36.7%) / NOT POTENTIAL CHANGE (201, 63.3%)	

have to identify punctuation mark words and replace them with symbols. The dictation tools detect the special words as commands and execute specific actions over each command. Users of these tools have learned over the years the available commands of each tool before using it. Although the commands nowadays usually take into account minor variants, they are not usually used for complex instructions due to their main transcription function. Mechanisms that switch from merely transcribing text and executing word-specific commands to incorporate in-context dialog with the assistant are required to have rich interactions.

6.1 Methods

The study of how users would interact with an interface that addresses document comments management in real settings requires the collection of real documents and the implementation of tools in the workplace. In this section, we explain the processes from documents data collection to the collection of interactions of participants in the field study.

6.1.1 Scenarios

The interaction over documents with comments is not the same for different types of comments. In order to identify what types of comments are present in documents, we collected documents publicly available on the Internet. We collected documents from CommonCrawl (Com, b) that range from 2013 to 2020. From the 107,885 .docx documents collected, only 1,313 of them were in English and had comments. A subsample of 100 documents were analyzed manually and identified three main types of comments: Modification, Information Exchange, and Social Communication. These categories resembles to previous work that identified for other domains (Dabbish et al., 2005). We then proceed to label the data via KarmaHub crowd workers (Kar). A random sample of 5,000 comments was labeled by three workers. The inter-rater reliability Cohen's kappa value was 0.65, indicating a substantial agreement. For every scenario identified in the manual inspection, we chose three samples. Table 6 shows the scenarios distribution.

[width=]interface.png

Figure 2: Document comment management user interface.

6.1.2 Interface

Now that we have real data, we need an editor interface capable of displaying the document and tracking the user interactions. Instead of using a traditional desktop editor to display the documents, we developed a web-based editor. This decision was based on the challenges associated with conducting crowd-sourced field studies on offline platforms. The editor was built using the CKEditor (WYS), a JavaScript library that includes the most common editor functions including document commenting. Figure 2 shows the different user interface elements. The user interface has three main sections: instructions, editor, and commands sidebar. The instructions explain the sequence of actions performed by the participant. The editor include a top bar from where the participants can change the format. The text to which the comment was assigned was highlighted in yellow. The comment associated to the commend was displayed at the right side of the paragraph. The commands side bar is a collection of transcribed voice commands. The speech-to-text transcription was performed via Microsoft Cognitive Services (Cog). In case a voice command was wrongly transcribed, the participant had the capability to edit in the comment sidebar.

Table 2: Intents and sub-intents

Category	Description	Example
MODIFICATION	The comment is a request for change, a commitment to making a change, or an acknowledgment of a change that was already performed.	Please write the answer in your own words.
MODIFICATION REQUESTED	Asking for a change.	I would add it as context for the pre-sales resource (in pink text).
CONTENT MODIFICATION	The modification is related to the content.	This could be rephrased to something like 'Once a study guide is available, all test candidates will be notified'
FORMAT MODIFICATION	The modification requires a change in formatting.	Should be centered throughout the doc
EXPLICIT	The things to be changed are explicitly defined in the comment.	We should remove this part of the statement.
NOT EXPLICIT	The exact changes are not explicitly mentioned within the comment.	Rephrase this bit
ADD	The comment is related to adding something.	3rd party, I assume? Please add to terminology table in section 1.2.
CHANGE	The comment is related to updating or replacing something.	Perhaps the criteria should be 'interchangeable in ALL context'
DELETE	The comment is related to removing something.	This section goes away since the content will be part of the VM.
EXECUTION	The reviewers inform the author of a change already performed or a promise to perform a task.	I added a few words to hopefully make it clearer.
DONE	It is informing that a change was made.	Added here
PROMISE	It is stating that a change will be made.	Sounds good I will start changing that everything
INFORMATION EXCHANGE	Comments that lead to exchange, analyze, verify, ask, request, or provide information.	What is the current process?
INFORMATION PROVIDED	Gives some context or provides some references.	See second paragraph here https://en.wikipedia.org/wiki/Regulatory
INFORMATION REQUESTED	Asks a question, clarify some content, or to validate something.	When and what should this notification communicate to the user?
CONTEXT	The reviewer supplies some contextual information.	The first version of the container images should be generated and ready before the MTP starts.
REFERENCE	The reviewer supplies references for reviewing.	See CT section for further issues.
ASKING DETAILS	The reviewer asks questions to retrieve more information.	Who gets this code?
REQUESTING CONFIRMATION	The reviewer asks the author to confirm something.	Is this the current matrix we generate and publish manually?
SOCIAL COMMUNICATION	Comments that provide feedback, acknowledge a comment, set communication beyond the document, or are part of a conversation that is not related to a change.	I think this is a good point.
ACKNOWLEDGMENT	The author is acknowledging a comment from a reviewer.	I see.
DISCUSSION	The comment is part of a conversation.	I'm glad there is an ongoing discussion
FEEDBACK	The reviewer gives feedback to the author.	Great start to this unit.
CONTENT RELATED	The comment is related to the content.	Providing a basic statement of why we're prioritizing these over others will help us negotiate when folks come to us with requests outside of this scope.
THREAD RELATED	The comment is related to the comment thread.	Feel free to add/edit to ensure this point is highlighted throughout the doc.
POTENTIAL CHANGE	After addressing it, it may lead to a change in the document.	Who is ́he ́?
NOT POTENTIAL CHANGE	It does not cause any change in the document after addressing it.	shared!

Table 3: Comparing F1 scores over the main level.

	LR	SVM	RoBERTa	DeBERTa	BART
Modification	0.75	0.74	**0.85**	0.84	**0.85**
Information Exchange	0.81	0.81	0.80	0.81	**0.82**
Social Communication	0.45	0.43	**0.69**	0.67	0.68
All	0.76	0.75	**0.82**	**0.82**	**0.82**

Table 4: Classification F1 results of the main level comparing sentence, comment, and their context.

	LR	SVM	RoBERTa	DeBERTa	BART
Sentences only	0.72	0.70	0.77	0.76	0.76
Sen. + Selected text	0.68	0.65	0.70	0.74	0.71
Sen. + Paragraph text	0.76	0.67	0.75	0.77	0.76
Sen. + Thread text	0.67	0.63	0.74	0.74	0.76
Comments only	0.69	0.68	0.80	0.79	0.81
Com. + Selected text	0.76	0.70	0.78	0.78	0.79
Com. + Paragraph text	0.73	0.73	0.73	0.81	0.79
Com. + Thread text	0.75	0.73	0.81	0.79	0.79
Sentences and Comments	0.75	0.75	**0.82**	**0.82**	**0.82**
Sen. & Com. + Selected text	0.78	0.75	0.77	0.79	0.80
Sen. & Com. + Paragraph text	**0.79**	**0.76**	0.80	0.79	0.79
Sen. & Com. + Thread text	0.74	0.75	0.81	0.80	0.80

6.1.3 Field Study

We conducted a crowd-sourced field study on KarmaHub. We iterated the instructions with the crowd-sourcing provider on three pilots to verify that the goals of the task were understood. We asked 50 participants to complete six scenarios each. We got three samples per scenario. Participants were asked to give a voice command first and then execute it in the interface. We collected voice samples and telemetry samples of each interaction. We paid workers 1.2 USD per scenario, considering an average time of 6 minutes per task and considering a wage of 12.0 USD.

6.2 Results

6.2.1 Text Commands Analysis

Table 7 shows metrics of how voice commands are composed. We found that most of the commands are short, and the mean range from 12 to 15 words across comment types. We detected that some of the words used in the commands were part of the contextual information. We define contextual information to text present in the comment, selected text, paragraph, or the task instructions. From the contextual content, the words in the comment were used more often (up to 23% of the words in the command text.) Most of the words (from 62% to 72%) were unique are were not present in the context.

Table 5: Comparing F1 scores over the main level intents and level one sub intents.

	LR	SVM	RoBERTa	DeBERTa	BART
Modification - Request	0.73	0.72	**0.75**	**0.75**	**0.75**
Modification - Execution	0.66	0.48	**0.79**	**0.79**	**0.79**
Info. Exch. - Request	0.64	0.61	0.80	0.79	**0.82**
Info. Exch. - Provide	0.71	0.70	0.76	0.75	**0.77**
Social Com. - Feedback	0.44	0.28	0.57	**0.62**	0.53
Social Com. - Acknow.	0.50	0.50	0.22	0.18	**0.80**
Social Com. - Discuss.	0.18	0.13	0.29	**0.32**	0.21
All	0.68	0.66	0.74	0.74	**0.75**

Table ?? shows the top ten trigrams detected on each type of comment. We can see that the most common trigrams correspond to phrases that were used to handle the comment box than phrases used to perform the requested edits.

6.2.2 Voice Commands Analysis

Table 8 shows the duration in seconds of each voice command. The voice commands range from 5 to 7 seconds, the median.

6.2.3 Telemetry Analysis

Table 9 shows the metrics obtained by analyzing the user actions in the experimentation platform. We can observe that participants spent a median between 7 to 20 seconds across conditions. Participants selected more text than the text that was typed. Not all the participants interacted with the comment box, the scenario with more interaction was social communication with 26%.

6.2.4 Multi-Modal Analysis

We identified that often the execution of the command took longer than the time to say the command; it ranges from 1 to 15 seconds. The number of selected words was longer than the words dictated by the users; this can be explained because of the use of ranges in the voice commands. Often users mentioned the first and end words of a sentence to mark the position from where to highlight a text.

6.2.5 Qualitative Analysis

After the command collection, the commands were separated by edition commands and comment management commands. We can identify how the assistant is impersonated, most participants were respectful by saying please before the commands i.e. "Please remove the text starting from [...]," "Please remove the text [...]." Some other users did not mention that they wanted to delete or resolve a comment; they only said, "Done." We identified some participants that delegated some tasks to the agent instead of retrieving and dictating manually "Please add the two journal titles that the co-author is asking."

Table 6: Scenarios

Scenarios	Category	Description	Example
1-5	MODIFICATION & REQUESTED & CONTENT & EXPLICIT & ADD	Comment requesting an explicit addition to the document	please insert "and the projects added or retired" between "baseline" and "beyond"
6-10	MODIFICATION & REQUESTED & CONTENT & EXPLICIT & CHANGE	Comment requesting an explicit change to the document	Change UNIT PRICE to LUMP SUM if appropriate.
11-15	MODIFICATION & REQUESTED & CONTENT & EXPLICIT & DELETE	Comment requesting a deletion in the document	Delete all document reference red or yellow highlighted text.
16-20	MODIFICATION & REQUESTED & CONTENT & NOT EXPLICIT & ADD	Comment suggesting something that implied the addition of content	Type an introductory sentence to this section of the report.
21-25	MODIFICATION & REQUESTED & CONTENT & NOT EXPLICIT & CHANGE	Comment with a suggestion that can derive to a change in the document	Not clear… please rephrase.
26-30	MODIFICATION & REQUESTED & CONTENT & NOT EXPLICIT & DELETE	Comment that suggests that something in the document is not required	Delete what is not applicable
31-35	MODIFICATION & REQUESTED & FORMAT & ADD	Comment that asks to add formatting	All URLs should be live links for the convenience of the reader.
36-40	MODIFICATION & REQUESTED & FORMAT & CHANGE	Comment that requests a change in the format	Should be in bold
41-45	MODIFICATION & REQUESTED & FORMAT & DELETE	Comment that asks to remove some formatting	You should not use bold for the title of your thesis/dissertation
46-50	MODIFICATION & EXECUTION & DONE	Comment that confirms that something was done	Changed from 6 grades per nine weeks to 10
51-55	MODIFICATION & EXECUTION & PROMISE	Comment that commits the author to perform a change	As you allowed, I will delete this text. Fully agreed.
56-60	INFORMATION EXCHANGE & PROVIDED CONTEXT	Comment that adds context to the select text in the document	Delivery of all deliverables required by the contract is usually a key requirement for revenue recognition.
61-65	INFORMATION EXCHANGE & PROVIDED REFERENCE	Comment that adds references to the text	See my previous comments on the Team discussion board
66-70	INFORMATION EXCHANGE & REQUESTED & ASKING DETAILS	Open question to the author	What is the border after this paragraph for? Is that a new subsection?
71-75	INFORMATION EXCHANGE & REQUESTED & REQUESTING CONFIRMATION	Question that requires the author to confirm something	I added this; does that make sense to include as a step?
76-80	SOCIAL COMMUNICATION & ACKNOWLEDGMENT	Comment that acknowledges that was read	Thank you for completing
81-85	SOCIAL COMMUNICATION & DISCUSSION & CONTENT	Comment that is part of a discussion that talk about the content	Further work on this to be discussed at the next meeting of AHIEC
86-90	SOCIAL COMMUNICATION & DISCUSSION & THREAD	Comment that is part of a discussion and is related to the thread	Same as above…
91-95	SOCIAL COMMUNICATION & FEEDBACK & CONTENT	Comment that provides feedback about the content	Good summary of what you found
96-100	SOCIAL COMMUNICATION & FEEDBACK THREAD	Comment that provides feedback to a comment in a thread	I am glad you folks are addressing these topics. These will be very helpful.

Table 7: Insights from text commands.

	Modifi.	Inf. Exch.	Soc. Com.
Words length (mean)	15	13	12
Chars length (mean)	88	84	67
Words overlap in comment	22%	23%	16%
Words overlap in selection	10%	4%	6%
Words overlap in paragraph	3%	3%	2%
Words overlap in instructions	11%	12%	9%
Unique words in the command	62%	65%	72%

Modification	Information Exchange	Social Comm.
delete the comment	delete the comment	delete the comment
no action needed	no action needed	no action needed
the comment please	thank you for	the comment no
the highlighted text	you for your	comment no action
task completed Delete	to user one	the highlighted text
the selected text	the comment no	action needed delete
end of the	comment no action	I have not
completed delete the	comment thank you	needed delete the
comment no action	reply to user	have not argued
HTTP colon forward	end of the	Thank you for

7 Discussion

The understanding of how users interact with voice interfaces for comment management can enable the development of smart assistants in the workplace. In this section, we discuss the results we observed in our field study and their potential applications.

7.1 Patterns in Voice Commanding

The complexity of resolving comments via voice relies upon the multi-actor nature of the task. A virtual assistant that mediates the communication between the authors has to understand the context of to whom the conversation is directed. The analysis identified commands that were related to editing the document and managing the comments.

Most of the edition commands follow the following structure: (1) Navigation Command; these were commands that place the cursor or identify the text to be formatted, deleted, or replaced (i.e., "At the end of the passage [...]", "[...] after the word [...]"); (2) Action Command, referees to a command that triggers an action such as format, add, replace, or delete part of the content (i.e., "Please delete the text [...]", "Insert the word [...]"); (3) Parameter Command, this works as the parameter of the performed action (i.e., "Replace the highlighted text with Dr. John Smith", "please use the word reps instead of representatives").

The comment management commands had low variability in the structure; we identified this common structure: (1) Action Command, a

Table 8: Insights from audio commands.

	Modif.	Inf. Exch.	Soc. Com.
Audio in seconds (mean)	6	5	7

Table 9: Insights from audio commands.

	Modif.	Inf. Exch.	Soc. Com.
Time performing changes (mean)	7	20	9
Number of selected words (mean)	22	16	13
Number of typed words (mean)	4	8	10
Interactions with the comment (%)	22%	28%	26%

request for deleting, replying, or marking the comment as done; (2) Dictation, when the action was "reply," then users started dictating the text to reply with.

7.2 Automatic Comment Management

The findings of this work can help platform designers to enable assistants in the text editors. From our results, we can observe that the time spent in dictation and in actually performing the task was similar. The main goal of those tools might not be to improve productivity but to offer hands-free solutions to manage collaborative documents. Tools can also help users triage their comments depending on the type of comment. The data can also be used to infer in which cases the users prefer to delete or to keep the comment.

7.3 Limitations and Future Work

The field study was conducted with crowd workers asked to resolve comments in documents that were not of their authorship and with comments left by strangers. The behavior of users that own the document and collaborate with people they know might differ the results. The participants did not work in a common text editor; this might cause a delay in their executions due to the lack of familiarity with the tool. Future work can conduct experiments in common text editors and with real teams to identify differences in the results.

Automatically handling comments can help people with visual impairment; however, the sample did not include that population, and it might not extrapolate. Future work can explore how people with visual impairments commonly interact with text editors and how they expect to manage document comments.

Our work focuses on the analysis of patterns in voice commands but does no further in the predictive analysis of the data. Future work can explore machine learning approaches that can automate tasks such as auto-completion, predicting when a comment is going to be resolved and other approaches that can push towards comment automation.

8 Conclusion

This work shed light on the required steps to automate document comment management. We explore how people interact with documents with comments. We first understand the different uses of comments in documents by analyzing public documents. We identified comments related to Modification, Information Exchange, and Social Communication. A sample of each category is presented to participants in a field study. We developed a platform that mimics a regular editor but with audio and activity tracking enabled. The participants were asked to provide voice commands and execute them manually to map the telemetry with commands. We identified the main commands used while interacting with the tool via voice, as well as the time spent on resolving each type of comment. We aim that the findings of this work can empower tools to support document comments management.

References

Cognitive services—apis for ai solutions | microsoft azure. https://azure.microsoft.com/en-us/services/cognitive-services/. (Accessed on 01/08/2022).

a. Common crawl. https://commoncrawl.org/. (Accessed on 09/15/2020).

b. Common crawl. https://commoncrawl.org/. (Accessed on 01/08/2022).

Karmahub | contact center | data | finance & banking services. https://www.mykarmahub.com/. (Accessed on 01/08/2022).

Wysiwyg html editor with collaborative rich text editing. https://ckeditor.com/. (Accessed on 01/08/2022).

Hanan Aldarmaki, Asad Ullah, and Nazar Zaki. 2021. Unsupervised automatic speech recognition: A review. *arXiv preprint arXiv:2106.04897*.

Madison Anzelc, Craig G Burkhart, and Craig N Burkhart. 2021. Can artificial intelligence technology replace human scribes?

Xuankai Chang, Wangyou Zhang, Yanmin Qian, Jonathan Le Roux, and Shinji Watanabe. 2020. End-to-end multi-speaker speech recognition with transformer. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6134–6138. IEEE.

Elizabeth F Churchill, Jonathan Trevor, Sara Bly, Les Nelson, and Davor Cubranic. 2000. Anchored conversations: chatting in the context of a document. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 454–461.

Laura A Dabbish, Robert E Kraut, Susan Fussell, and Sara Kiesler. 2005. Understanding email use: predicting action on a message. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 691–700.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Anil Kumar Gupta, Ankita Kumari, and Rachna Somkunwar. Automatic speech recognition and transliteration.

Morris Halle and Kenneth Stevens. 1962. Speech recognition: A model and a program for research. *IRE transactions on information theory*, 8(2):155–159.

N Ratna Kanth and S Saraswathi. 2015. Efficient speech emotion recognition using binary support vector machines & multiclass svm. In *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pages 1–6. IEEE.

Asad Khattak, Anam Habib, Muhammad Zubair Asghar, Fazli Subhan, Imran Razzak, and Ammara Habib. 2021. Applying deep neural networks for user intention identification. *Soft Computing*, 25(3):2191–2220.

Daniel Lewis. 2006. What is web 2.0? *XRDS: Crossroads, The ACM Magazine for Students*, 13(1):3–3.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Gordon E Peterson. 1961. Automatic speech recognition procedures. *Language and Speech*, 4(4):200–219.

Roberto Pieraccini and ICSI Director. 2012. From audrey to siri. *Is speech recognition a solved problem*, 23.

V Radha and C Vimala. 2012. A review on speech recognition challenges and approaches. *doaj. org*, 2(1):1–7.

D Raj Reddy. 1976. Speech recognition by machine: A review. *Proceedings of the IEEE*, 64(4):501–531.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wei Wang, Saghar Hosseini, Ahmed Hassan Awadallah, Paul N Bennett, and Chris Quirk. 2019. Context-aware intent identification in email conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 585–594.

Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2017. Identifying semantic edit intentions from revisions in wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2000–2010.