Education Tech Research Dev (2022) 70:1755–1779
https://doi.org/10.1007/s11423-022-10137-5

DEVELOPMENT ARTICLE

Exploring collaborative caption editing to augment video-based learning

Bhavya Bhavya[1] · Si Chen[2] · Zhilin Zhang[1] · Wenting Li[1] · Chengxiang Zhai[1] · Lawrence Angrave[1] · Yun Huang[2]

Accepted: 23 June 2022 / Published online: 15 July 2022

Abstract
Captions play a major role in making educational videos accessible to all and are known to benefit a wide range of learners. However, many educational videos either do not have captions or have inaccurate captions. Prior work has shown the benefits of using crowdsourcing to obtain accurate captions in a cost-efficient way, though there is a lack of understanding of how learners edit captions of educational videos either individually or collaboratively. In this work, we conducted a user study where 58 learners (in a course of 387 learners) participated in the editing of captions in 89 lecture videos that were generated by Automatic Speech Recognition (ASR) technologies. For each video, different learners conducted two rounds of editing. Based on editing logs, we created a taxonomy of errors in educational video captions (e.g., Discipline-Specific, General, Equations). From the interviews, we identified individual and collaborative error editing strategies. We then further demonstrated the feasibility of applying machine learning models to assist learners in editing. Our work provides practical implications for advancing video-based learning and for educational video caption editing.

Keywords Lecture video caption editing · Caption transcription · Collaborative editing · Technology-assisted editing

Introduction

As the popularity of video-based learning (e.g., MOOCs) continues to grow, it is crucial to provide accessible online videos to all, including but not limited to, users who are non-native speakers and learners with disabilities. The COVID-19 pandemic has necessitated a rapid shift to the use of online educational videos, increasing the urgency of providing accessible online educational videos (McCarron, 2021). Prior research has shown that accurate captions play an

✉ Bhavya Bhavya
bhavya2@illinois.edu

[1] Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[2] School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA

important role in making lecture videos accessible (Cross et al., 2019) , especially for learners who are Deaf or Hard of Hearing (DHH). Further, following the principles of Universal Design Learning (UDL), captions are found to be beneficial to a wide range of learners (Clossen, 2014) that also includes learners with other disabilities (e.g., ADHD, dyslexia) and learners who prefer to learn by reading or searching transcriptions.

However, creating captions is challenging. Commercial captioning services have a slow turnaround (requiring a business day or longer) and are expensive (approximately $1 per audio-minute) (Wald, 2013), while centralized university services have capacity limits which can only take on a small fraction of the total number of content hours generated by a university each each day. Alternatively, captions generated by Automatic Speech Recognition (ASR) algorithms are cheaper (e.g., Microsoft and Google cloud ASR services are approximately $1 per audio-hour) but have been shown to be too inaccurate to be used exclusively for learning with educational videos (Parton, 2016). Generating domain-specific captions, e.g. captions for Science, Technology, Engineering, and Math (STEM) courses, presents additional challenges due to the use of scientific nomenclature and diverse instructors with accents (Ranchal et al., 2013; Lewis, 2021).

Prior work has achieved significant success in using crowdsourcing to edit ASR generated captions to lower the cost of obtaining high-quality captions (Huang et al., 2017; Culbertson et al., 2017). However, how learners edit captions of domain-specific educational videos is not yet investigated. For example, what are the challenges of captioning domain-specific lecture videos for individual learners and for groups of learners when they collaboratively edit captions? Gaining an empirical understanding and addressing the above questions can advance algorithms that improve captioning services for educational videos and suggest opportunities for human-in-the loop captioning.

In this paper, we deployed a system for collaborative caption editing at a large enrollment ($N = 387$) course at a US university where 58 learners edited 89 lecture videos in a STEM course. The captions were edited in two rounds by learners. Follow-up interviews were conducted with 18 of those learners to further understand the challenges, strategies used, and need for better support in editing STEM captions. We conducted qualitative analysis on the interview results, performed quantitative system log analyses for understanding editing behavior specific to STEM captions, and examined the results using the widely-adopted "Find-Fix-Verify" crowdsourcing paradigm (Bernstein et al., 2010).

Our findings make the following contributions to the Educational Technology community. First, we developed a taxonomy for caption edits for educational videos, which also allowed us to gain new and empirical understandings of individual and collaborative editing behavior within different categories of edits. Second, we identified and categorized strategies for individual and collaborative caption editing specifically in the context of video-based learning. These strategies informed future design of educational video captioning systems. Third, inspired by learners' challenges and suggestions, we evaluated the feasibility of applying machine learning algorithms to better support captioning editing in educational videos. Additionally, we proposed new designs and discuss theoretical and practical implications of collaborative captioning service for video-based learning.

Related work

Caption generation for educational videos

Research has shown that to make the content of educational videos accessible to the widest audience possible, it is important to improve the readability of the text and captions of the videos (Cross et al., 2019). learners who are Deaf or Hard of Hearing rely on captions for access to video content. Further, following the principles of Universal Design for Learning (UDL), captions are found to be beneficial to a wide range of learners (Clossen, 2014). Yet many videos remain uncaptioned or have machine-generated captions with high error rates (Shiver & Wolfe, 2015).

For example, machine-generated captioning via open source models (e.g. Mozilla DeepSpeech) and commercial cloud-based Automatic Speech Recognition (ASR) services including cloud services by Microsoft, Amazon and Google and start-up companies (e.g., Otter.ai) provide free or low-cost captioning. However, they do not meet the Americans with Disabilities Act accuracy goal of 99% Word Accuracy Rate for publicly available video content (Klein, 2021). A 2020 analysis found that Google's API outperformed IBM Watson and Facebook's Wit.ai ASR, but recorded an average word error rate (WER) of approximately 9% (Filippidou & Moussiades, 2020). Practically a WER of 9% is a significant barrier to efficient and accurate learning; a learner would be attempting to learn with mistakes and inaccurate statements that occur in most sentences. Compared to general videos, captioning educational videos correctly is even more challenging as it requires more domain knowledge. For example, automatically generating video captions for STEM courses presents unique challenges due to substantial scientific nomenclature and technical terminology (Ranchal et al., 2013).

Instead of using fully automatic methods, researchers have also conducted case studies where human editors correct errors to improve the readability of the lecture transcripts for learner use and enhance the accuracy of future speech recognition models (Ranchal et al., 2013; Valor Miró et al., 2014). Nevertheless, error correction was still found to be the most time-consuming task to create lecture transcripts (Ranchal et al., 2013); suggesting the need for a more scalable way to generate accurate transcripts.

Collaborative video captioning

Researchers have been studying crowd-sourcing video caption systems that harness both low-cost automatic generated caption and human intelligence in video caption creation. Huang et al. leveraged complementary contributions of different workers to design, implement and evaluate an efficient crowd-sourcing system for video captions—BandCaption (Huang et al., 2017) and Mahipal et al. created a similar system—ClassTranscribe (Mahipal et al., 2019) that was evaluated in several large enrollment (N > 200) computer science classes and allowed instructors to reward individual learners with credit for their crowdsourced edits. These systems combine automatic speech recognition with input from crowd workers to provide a cost-efficient captioning solution for online videos.

Several theoretical models for effective collaboration in crowdsourced systems have been proposed. The "Mark-Edit-Approve" model allows subsequent workers to edit earlier workers' edits to conduct crowd-captioning quality control (Huang et al., 2017). The "Find-Fix-Verify" model has been shown effective in various collaborative crowdsourcing systems, such as crowdsourced writing (Bernstein et al., 2010) and micro-task assignment (Bozzon et al., 2012). The Find stage, asks workers to identify patches that need more attention; the Fix stage recruits workers to revise an identified patch; the Verify stage performs quality control on revisions by recruiting workers to vote on others' work. We use this model to ground our findings.

There exist educational systems, such as ClassTranscribe (Ren et al., 2015) and ICS Videos (Deshpande et al., 2014), that can generate captions for lecture videos in a collaborative fashion with learners. Previous studies (Deshpande et al., 2014; Angrave et al., 2020a; Cross et al., 2014) have shown that such tools are effective and efficient for captioning lecture videos and have considerable value in educational practice (Angrave et al., 2020a, 2020b; Amos et al., 2021; Zhang etal, 2021; Zhang, 2021). Previous research has also shown that involving learners in fixing captions in foreign language educational videos does not impair learning and also helps reduce errors in the captions (Culbertson et al., 2017). However, despite the fact that these tools could be beneficial for educational purposes, none of these works studied how learners worked together in the experience.

To the best of our knowledge, ours is the first study to focus on learners' collective behaviors towards improving the automatically generated captions of lecture videos in an online class. By better understanding human editing of machine-generated information, we are also able to propose design insights on new ways for future systems, including deeper human-machine collaboration. This interactive process between machine learning and humans is called 'Interactive Machine Learning' (Amershi et al., 2014). One example of leveraging the human-edited ASR captions is to further train automatic error identification and correction models (Errattahi et al., 2018; Hrinchuk et al., 2020). However, the effectiveness of such models for automatically correcting captions of educational videos based on learner edits has not been studied before.

Specifically, in this paper, we addressed the following research questions:

RQ1 How do individual learners make edits to crowd-sourced captions?

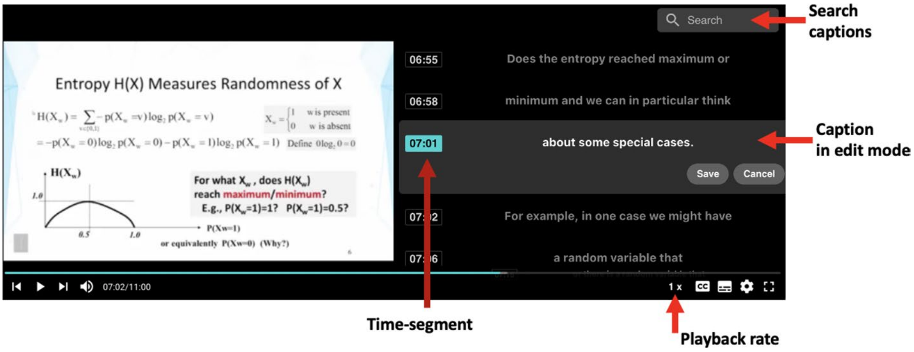
RQ2 How do learners collaborate with other learners in crowd-sourced caption editing?

RQ3 How can the system better support learners to conduct caption edits for educational videos?

Research method

In this section, we describe our study design and methodology. Overall, the case study method was adopted because it has been widely adopted for exploratory and descriptive educational research such as ours (Tellis, 1997; Hamilton & Corbett-Whittier, 2012) as it offers a framework for holistic investigation and understanding of complex social units (Merriam, 1985). We follow the suggested best practices to improve the credibility of the findings including triangulation using multiple data sources (Merriam, 1985).

Below, we describe the data collection process including the system used for editing ASR generated captions, the caption editing activity where 58 learners in a Text Mining course with 387 learners participated in editing captions of 89 out of 93 lecture videos with two editing rounds, and follow-up interviews conducted with 18 learners to understand their editing experience and behavior. Next, we describe the data analysis methods


Search
Entropy H(X) Measures Randomness of X
For what X_w, does H(X_w) reach maximum/minimum?
E.g., P(X_w=1)=1? P(X_w=1)=0.5?
07:02/11:00
1 x
Save
Cancel

06:55 Does the entropy reached maximum or

06:58 minimum and we can in particular think

07:01 about some special cases.

07:02 For example, in one case we might have

07:06 a random variable that

Search captions

Caption in edit mode

Time-segment

Playback rate

Fig. 1 Interface for editing captions

that included coding interview data, coding and statistical analysis of caption edit log data, and building machine learning models to support caption editing.

Data collection

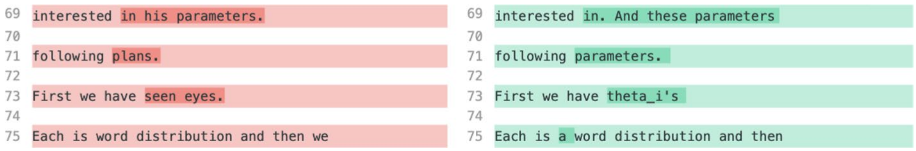
ClassTranscribe: a video-based learning system

ClassTranscribe (Mahipal et al., 2019; Angrave et al., 2020b) is an open-source web platform for delivery of educational online lecture videos. In this system, the lecture audio is initially transcribed using an automatic speech-to-text cloud service (Microsoft Azure Cognitive Services Speech-To-Text) at a cost of approximately $1 per audio-hour. The captions are indexed to enable keyword-based search. The system is built and deployed as a set of Docker containers on a Linux virtual machine with a Postgres SQL database. The database schema and design choices are published in (Mahipal et al., 2019). Source code is available at https://github.com/ClassTranscribe.

Figure 1 shows the system user interface for editing captions. The lecture captions are displayed on the right side of the video. Each caption is annotated with the corresponding video time-segment. By clicking on the time-segment, users can jump to that video moment. Clicking on a caption opens up the caption in the “edit mode”. After editing the caption, users can click on the “Save” button or press “return/enter" on their keyboard to save their edits. Any edits made are instantaneously reflected on the interface. Users can also search for keywords within captions across all course lecture videos. Captions can also be turned off. Other useful features include adjusting the playback rate, pausing/playing video, adjusting the video progress bar (seeking) etc. All user activities on the system, including searching, watching a video, seeking, editing a caption are logged in the SQL database for later analysis.

Caption editing activity design

In Fall 2020, the lecture videos of an online senior-level computer science course on Text Mining and Analytics at a large public university in the US were uploaded onto the ClassTranscribe system. Learners in the course were provided with an opportunity to participate in an extra-credit activity to fix errors in captions in lecture videos. There were a



```
interested in his parameters.          interested in. And these parameters

following plans.                       following parameters.

First we have seen eyes.               First we have theta_i's

Each is word distribution and then we  Each is a word distribution and then
```

Fig. 2 Sample captions *before* and *after* an edit

total of 93 lecture videos in the course. Each video was on average 12 minutes long. Each lecture video transcript had about 1500 words.

The caption editing activity was divided into two sequential tasks. The task of *editor one* was to take a first pass at correcting errors in captions. After *editor one* completed their task, *editor two* then reviewed the captions to fix any remaining errors. Learners were unaware about the exact edits made on the videos by other learners.

The ClassTranscribe system was released to learners towards the end of the course only for this activity. Prior to that, learners used Coursera[1] to watch the online course lectures. Interested learners were provided with an opportunity to sign up as *editor one* and *editor two* for one lecture video each to get 1% extra-credit. To enable participation from many learners, each learner could sign up for a maximum two lecture videos each as *editor one* and *editor two* for a total of 2% extra-credit. Learners were given two weeks to complete the *editor one* task first. The following two weeks were for completing the *editor two* task. Other extra-edit activities were also released simultaneously to give all learners a fair chance at receiving extra credits. Besides, we did not set any minimum number of edits to get the extra credit.

For every caption edit made on the system by learners, the log captured who made the edit, the time of edit, the caption before and after the edit, the corresponding lecture video time-segment, and the lecture name. Figure 2 shows four captions before (left) and after (right) edits. The edited words are highlighted. As can be seen, each caption is not necessarily a complete sentence because captions are segmented based on the corresponding video time-segments. Further, multiple words could be edited and logged within a single caption-level edit. 58 learners edited 89 out of 93 lecture videos; learners signed-up for editing all the 93 videos but ultimately did not edit 4 videos. Further details about the log statistics are presented in the findings throughout the rest of this paper.

Follow-up interview

In Spring 2021, learners who had previously edited captions were recruited by email to participate in an online semi-structured interview. All interviews were conducted on Zoom and were recorded with consent from participants. The study was approved by the University IRB.

Each interviewee was paid at $20/hour (pro-rated) and interviews were approximately 45 minutes on average. Table 1 summarizes the demographics of the of the 18 interviewees. None of the interview participants reported having any physical or mental chronic conditions that would prevent them from understanding the speech in lecture videos.

[1] https://www.coursera.org/.

Table 1 Interviewee demographics

PID	Gender	Age	Race	Major	Program	English proficiency	V. edits
P1	F	18–24	A	CS & Stats	UG	NN	No
P2	M	35–44	W	CS	DSG	N	Yes
P3	F	18–24	A	CS	DSG	NN	No
P4	M	25–34	A	CS	DSG	N	No
P5	F	18–24	A	CP	UG	NN	No
P6	F	18–24	A	CS	G	NN	No
P7	F	18–24	A	CS	G	NN	No
P8	M	35–44	A	CS	DSG	NN	Yes
P9	M	25–34	A	CS	DSG	N	No
P10	M	18–24	A	CS	UG	NN	No
P11	F	18–24	A	CE	UG	NN	No
P12	M	18–24	A	CEE	UG	NN	No
P13	M	25–34	A	CS	DSG	NN	No
P14	F	18–24	A	CS	DSG	NN	No
P15	F	18–24	A	AE	UG	N	No
P16	–	25–34	A	CS	DSG	NN	No
P17	M	35–44	W	CS	DSG	NN	Yes
P18	M	25–34	B	CS	DSG	N	No

There were eight Females (F), nine Males (M) and one who preferred not to disclose their gender. Fifteen participants identified as Asian or Asian American (A), two as White (W), and one as Black (B). Thirteen participants were from Computer Science (CS). Other majors included Computer Engineering (CE), Aerospace Engineering (AE), Cognitive Psychology (CP), Civil and Environmental Engineering (CEE), and Computer Science & Statistics. Ten participants were working professionals in the Data Science graduate degree program (DSG), six other graduate learners (G) and two undergraduates (UG). Thirteen participants were non-native English speakers but reported no problems in understanding or speaking English (NN) and five native English speakers (N). Three participants performed some voluntary additional edits ("V. Edits") without officially signing up in the caption editing activity

The major questions asked in the interview were as follows: why and how participants use captions for learning with educational videos; could they provide examples of different errors they noticed in the captions and explain how they impact their learning; explain how and why they edited specific errors in captions; if/why they chose to ignore certain errors; could they explain the difference between being *editor one* and *editor two* in terms of their behavior, experience, effort or process of editing; could they explain the impact of the editing activity on their learning; could they explain their motivation to make edits; and provide suggestions to improve the system to better support caption editing and improve their overall experience. To facilitate the discussion related to their editing process and experience, we showed the interviewee their editing log (2) comparing captions before and after their edits and probed them to explain any interesting or important edits.

Data analysis

Coding of interview data

We followed previous work to first establish properties of what participants said without relying on existing theories (open coding), proceeded to identify relationships among the codes (axial coding), and conducted a thematic analysis on the interview data in a similar process described in (Dye et al., 2019; Dym et al., 2019). First, two authors familiarized themselves with the data by reading the transcripts carefully. Two authors then performed open coding with four participants' transcripts independently and then met to discuss and compare their codes. They then discussed discrepancies and revised and expanded the existing categories until they reached an agreement of three main themes—'Individual Editing Strategy', 'Collective Editing Strategy', 'Suggestions to Improve Editing Experience'. Finally, the first author coded the remaining data through an iterative process, in which she met with the second author regularly to discuss the codes and iterate on the findings

Caption edits log analysis

We analyzed 10, 378 rows of word-level edits generated by the 58 learners who conducted caption editing. To better understand what types of edits were made by the learners, two authors manually coded and analyzed word-level edit logs. Multiple edits may occur in one caption, so we transformed our editing log from caption-level to word-level. In addition to regular English word edits (e.g., *world* → *word*), we also analyzed edits made to punctuation characters (10% of total edits), numbers and Greek letters (4% of total edits), and capitalization (6% of total edits). We refer to all such edits as word-level edits, which includes 1904 unique words. A word could be in multiple categories/subcategories, and was labeled based on its context and usage in the lecture. For example *Map* in the context of *Map Reduce algorithm* is labelled as Discipline-Specific edit, whereas in the context of *Map out* is labelled as General Edit.

The first-round inter-rater agreement was 76% (Edits)–94% (Discipline-Specific Edits)–96% (Symbol Edits). The two authors discussed finalizing all disagreed cases one by one by closely reading "before" and "after" edited captions in sentences, referring to original video clips.

To compare various learners' editing activities and edit types, we conducted non-parametric Wilcoxon Statistical Tests (Wilcoxon, 1992).

Machine learning model for supporting caption editing

The rich editing log data collected from students could potentially be used to train machine learning models to support learners to conduct caption edits. We first introduce some terminology to facilitate the discussion of such machine learning models. Our log data captures caption-level edits. Consider the unedited caption c_{bef} "*This lecture will look"* that was edited to c_{aft} "*this lecture, we will*". There are two new words inserted or substituted in c_{aft}, i.e., "*this*", "*we*". We refer to them as w_{aft}. Each w_{aft} is manually

Table 2 Taxonomy of caption edits

Level I	Discipline-specific				Gen.	Typo
Level II	Symbol			Non-Sym.		
Level III	Eq.	Ab.	Sgl. Sym.	DS. Ph.		
Example	cedar → theta	PSA → PLSA	D1 → d1	tax → text	the → like “the”	waiting → weighing

Gen. = General, Non-Sym. = Non-Symbol; Eq. = Equation, Ab. = Abbreviation, Sgl. Sym. = Single Symbol; DS. Ph. = Discipline-Specific Phrases

coded as described in “Caption edits log analysis”. We train models to predict the category of w_{aft} to assist learners while editing.

To construct train and test datasets, we chronologically split the entire manually coded data into train-test splits based on the timestamp of edits. We chose this way of constructing the train and test sets to mimic a real scenario in a course where the initial edits were made by learners could be used for training. Prior text classification work has also constructed the train/test datasets in a similar chronological fashion, e.g., (Klimt & Yang, 2004). We used a $80-20\%$ train-test data split. After removing Typos, there were 8k and 2k samples (w_{aft}) in the train and test sets respectively. As discussed in Section “A taxonomy of caption edits in educational videos” , capitalization, punctuation, etc. are important for this task, so lower-casing and other standard text pre-processing steps were not applied and the raw data was used instead.

As our main objective was to test the feasibility of training classifiers using the collected data and not necessarily find or design the best classifier for these tasks, we employed standard text classification algorithms Logistic Regression and Random Forest with standard bag-of-words features using unigram tf-idf word vectors (Aggarwal & Zhai, 2012). We also used class-weighting (Kotsiantis et al., 2006) to handle imbalanced class distribution that is discussed in the results section “Machine learning model for efficient edit verification”. Hyperparameters were tuned using grid search with a 5-fold cross validation on the training set. The baselines include two standard baselines often used in classification tasks. Majority baseline classifies every word as the majority class. Random baseline uniformly randomly selects a category to assign to each words.

Results

In this section, we describe the results of our data analyses to address the three research questions.

Individual learners' caption editing behavior (RQ1)

By analysing the editing log by edit type and analysing follow-up interviews, we generated deeper insights on different types of edits to answer RQ1: How do individual learners make edits to crowd-sourced captions? By coding the edit log, deeper insights on the edit types was obtained. Further, as described in the “Follow-up interview” section, during the interviews, interviewers went through each participants' editing log that highlighted captions

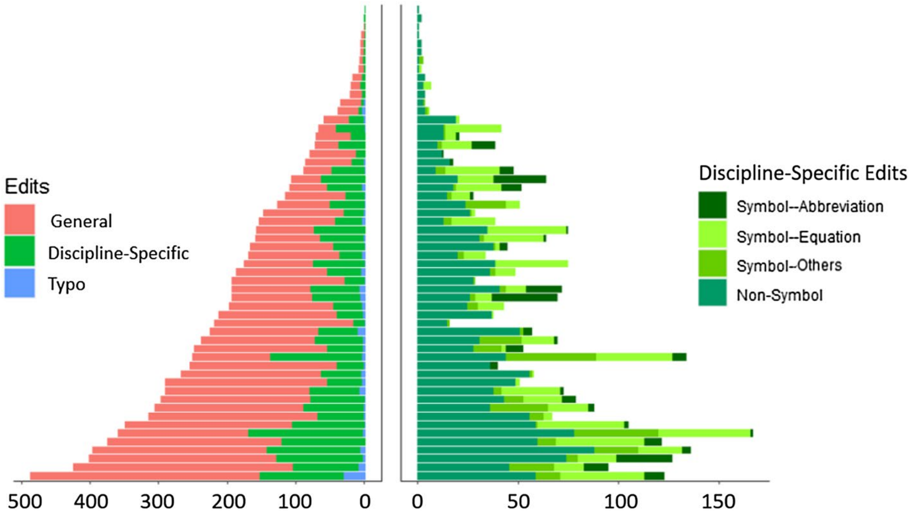


Fig. 3 Numbers of different types of edits made by 58 learners. Left: Edits: Discipline-Specific Edits, General Edits, Typo Edits. Right: Discipline-Specific Edits: Symbol Edits (Abbreviation, Equations, Single Symbol), Non-Symbol Edits. Discipline-Specific Edits on the left side and on the right side is of same value

before and after edits. This allowed interviewers to explore and understand the learners' editing goal, editing process and editing strategies using explicit editing examples. Then from interviews, we found all participants spontaneously followed the "Find—Fix—Verify" (Bernstein et al., 2010) steps in caption editing process by identifying errors in exiting captions, finding the correct way to edit errors, and verifying edits made. Various strategies were applied by different participants at individual steps. Below, we present different strategies identified in the three stages correspondingly.

A taxonomy of caption edits in educational videos

Table 2 shows the categories of edits we identified via coding. Levels I, II, and III indicate the category hierarchy. For example, *symbol edits* and *non-symbol edits* are sub-categories of *discipline-specific edits* . At Level III, *equation edits* are edits made to correct characters or punctuation in equations or formulae symbols; *abbreviation edits* are edits made to correct short forms, and *single symbol edits* are other edits that modify a single symbol specific to the course or discipline. Since these three types of edits were specific to symbols (i.e., a mark or character used as a conventional representation of an object, function, or process), these were grouped into *symbol edits* . *Discipline-Specific Phrase* edits are edits made to correct longer phrases (i.e., *non-symbols*). These are all sub-categories of edits specific to words about the subject or discipline (*discipline-specific edits*). In contrast, *general edits* are edits made to correct general english words or punctuation that are not related to the discipline.

Typo refers to a misspelled word with absent or additional letter(s) or letter sequence errors that arise while using keyboard for text input. The *typos* category was added due to obvious spelling errors noticed while reading captions; it does not indicate a mismatch between audio content and captions.

Table 3 Different types of edits and their edit frequencies

Type of edits	# of Edits/word	# Of edits/learner
Equation	7.1	11.2
Abbreviation	3.4	4.0
Single symbol	4.7	5.7
Discipline-specific phrases	3.3	20.8
General language	8.3	123.2
Typo	1	3

We acknowledge that the subcategories might be different for caption edits in other courses and disciplines and further study is needed to identify general taxonomies, e.g., *equation edits* might only be found in some STEM courses. However, our study is a first step towards this broad goal.

Frequency of different types of edits In total, 58 participants revised in total 10, 378 edits (Mean = 178.9, SD = 131.2). As shown in Fig. 3, there were 7143 *general edits* (Mean = 123.2, SD = 151.8), 3061 *discipline-specific edits* (Mean = 52.8, SD = 41.8), and 174 *typos* . There were significantly more *general edits* than *discipline-specific edits* (Wilcoxon test, W = 2333.5, p < .001). Within *discipline-specific edits* , there were 1853 *symbol edits* (Mean=31.9, SD = 22.1), 1209 *non-symbol edits* . There were 647 *equation edits* , 232 *abbreviation edits* , and 330 *single symbol edits* . Using a Wilcoxon test, there were significantly more *equation edits* than *abbreviation edits* (W = 2071, p < .03), *single symbol edits* (W = 1119.5, p < .001), and *non-symbol edits* (W = 432.5, p < .001).

The 10,378 edits include 1904 unique words. Such results indicate that there were repetitive edited word. Here, repetitive edits mean after-edit word is the same; it does not mean words before editing are the same. In Table 3, we printed frequency of edits per unique word on the second column. *General edits* included more repetitive edits than *discipline-specific edits* , where 864 unique *general* words were in total edited 7142 times (Mean = 8.3. SD = 34.0). For example, missing "," was added 547 times. Within *discipline-specific edits* , *symbol edits* and particularly *equation edits* were most likely to include repetitive edits. For example, "*theta*" was edited 130 times, such as from "*P of Cedar one is the probability of*" → "*P of theta one is the probability of*". There were no repetitive *typos* , which means there were 174 unique word-level typos.

In sum, learners' caption editing log shows that there were significantly more *general edits* than *discipline-specific edits* Within *symbol edits* , there were significantly more *equation edits* than *abbreviation edits* , *single symbol edits* , and *non-symbol edits* .

Individual editing strategies

Interview data revealed that the main goal for editing captions was to improve the accuracy of captions. Participants described three steps to improve accuracy—finding errors, fixing errors, and verifying errors—and the strategies employed at each step are discussed below, along with the factors that impacted those strategies. The agreed editing criteria was not to represent exactly what the lecturer spoke, but to seek a balance between accuracy and efficiency. Other secondary goals included to learn content better (e.g. better prepare for exams), and improve confidence in caption editing. Learners used different strategies across different steps to improve accuracy and increase efficiency.

Identifying errors Strategies included guessing; comparing audio, slides, and captions: Our participants perceived that "noticing error" was a rather easy task than making actual edits and "providing the correct answer". To improve efficiency, some learners directly identified the error by looking at captions and identifying words that "were not supposed to appear in the context, such as the word *Opera*" (P12). Another mainly used approach to improve accuracy was watching the video and simultaneously skimming captions. Under such an approach, they could notice any mismatch between audio and caption, slides and captions.

Error type impacted perceived likeliness of noticing errors: Different strategies were described in identifying errors in *discipline-specific* and *general* content. For *discipline-specific edits* - *non-symbol edits* were most likely to be noticed. One participant explained 'It does not require any prior understanding of what the last sentence was about, what term the professor is talking about. I can get easily identify it by glance "(P13). For example, "*tax retrieval* " → "*text retrieval*". General Edits were harder to notice.

Prior domain knowledge impacted likelihood of noticing domain-related errors: A majority of participants agreed that editing captions towards the end of the semester allowed them to have a general understanding of the course and be more familiar with domain-related terms. Namely, P11 thought it might be harder to capture Discipline-Specific errors when watching the video for the first time because the participant would be less familiar with them e.g., might not recognize "*organize*" should be "*tokenize*".

General usage of captions impacted likeness of noticing errors: Context of learning impacted learners' reliance on captions to access knowledge. For example, P15 said he sometimes just read the caption without opening the audio. Additionally, proficiency of languages, in our case English, impacted their reliance on captions. P6 said as a non-native speaker, she found reading easier than listening. She could not stop herself from looking at captions while taking classes in English, which would never happen when consuming videos of native languages (Chinese). In such cases, it was more likely for participants to notice errors in captions since they read them frequently.

Access to visual information impacted likeness of noticing errors: Slides provided learners with information-rich references, and they recognized errors in captions when there was a mismatch between caption terms and slide terms. But, having access to slides sometimes also made it harder to notice errors. For example, P13 said "I might auto-correct that (error) in mind by looking at the slides, missing such errors, but it needs to be fixed."

Editing errors Strategies included prioritizing edits; guessing; listening carefully; reading slides carefully: A widely used strategy to improve efficiency was to prioritize edits based on errors' impact on understanding course content. Many of our participants classified errors as major errors and minor errors based on perceived understanding and prioritized major errors, "it's not important to understand every single word that instructor says—only major concepts" (P11). Another strategy to improve accuracy was to replay a video clip multiple times and read slides closely. Another interesting way to improve the efficiency was to guess the correct word based on context of sentence and slides. Sometimes, they mentioned they prioritized readability over precise match between caption and video timestamp.

Edit type impacted editing effort: Domain-related words were perceived as major errors impacting learners' understanding the most, "*parody*" → "*paradigmatically*" is an example for change of meaning (P4). Minor errors were defined as errors that were unlikely to change meaning of a sentence and didn't impact the context of a sentence. Such errors were either ignored, or sometimes revised, or left for a second pass (P9, P10, P11). An example is capitalization indicating a start of sentence (P8, P13, P14, P17). Such errors were also refereed as "grammatical errors" by some participants (P4, P5, P7, P8).

According to interview results, *discipline-specific edits* were considered to be correcting major errors and of top priority to edit. Within *discipline-specific edits* , *equation edits* have a major impact on understanding course content and high priority edits. Participants understood that auto-generated captions were especially challenging in generating equation captions correctly. P12 said "they would need to use the correct presentation for numbers, especially Greek alphabet and punctuation that I don't think they even have." Some examples include "*H*(*X*|*Y*)"; "*I(X;Y)*"; "*theta*". *Abbreviation edits*, also described as "*jargon*" (P7), also had a major impact on understanding. It was especially confusing when misspelled in various ways, such as "*PLSA*" misspelled as "*PLA*", "*SL*", "*PLC*", etc. Another widely reported case was when system wrongly transcribed *discipline-specific* words into simple words and changed the context (e.g. "*And complete this magic.*" → "*and completely symmetric.*")

General usage of caption impacted perceived error importance: In many interviews, participants mentioned copying and pasting captions for note-taking. This required higher caption quality and they had less tolerance in caption errors (P5, P6, P7). Similarly, some participants mentioned errors impacted their usage of "caption searching" feature to better capture the concepts they were interested in reviewing (P15, P17).

Verifying edits Strategies included proof-reading; referring to outside sources; skipping errors; following transcription norms: During the interview, half participants felt confused or were uncertain while making edits and trying to provide the correct text. Most participants edited all errors that they noticed; a few ignored unsure errors and "left it for others" due to uncertainty (P9, P10, P13). To address uncertainty and improve accuracy, participants re-read captions or verified the correct spelling and formatting of domain-related terms and symbols with external trusted sources (P2, P3). Below, we describe the factors impacting confusion and uncertainty.

Edit type impacted editing confidence: Participants were more certain in editing *discipline-specific edits* compared to *general edits* . Participants reported that the corrected text of *discipline-specific* errors could often be found be on the slides, and they only needed to type the corrected word manually, though it could be tedious. For *general edits* , participants needed to closely re-listen to video clips multiple times and play at a slower speed (e.g., .75-times original speed) to identify the correct word; they made sure they were spelling it correctly. They also found such edits to be challenging for non-native English speakers (P3, P12, P15, P16). Some participants mentioned they used "guessing" to propose a correct word. Within *discipline-specific* , participants expressed less confidence in making *equation edits* as it required greater "cognitive overload" and they "struggle with formatting correctly" e.g., how to transcribe numbers and Greek numerals.

Prior domain knowledge impacted editing confidence and confusion: A majority of participants agreed that editing captions towards the end of the semester made them more confident when editing all kinds of errors. Five participants thought such domain knowledge could be especially helpful in correcting domains-related terms. As mentioned in prior sections, providing a correct answer required domain knowledge.

Participants felt confident editing captions when they had domain knowledge (P11) and would be less likely to "be confused about whether the word is something they don't know, causing unnecessary doubts for new learners..." (P12).

Familiarity with transcription task impacted editing confidence and confusion: An important factor that impacted learners' confidence and confusion with caption editing was familiarity with the transcription task. The participants described the transcription task as "subjective" and were unsure of the extent to which the transcription should record exactly what the lecturer said. For example, P9 said "I don't know if I should be 100% accurate, should I be typing down every word the professor say, or should I rephrase it in some way." Other participants reported hesitation on whether to include filler words and repeated words (e.g., "*um*") (P15, P2, P3, P7). Participants also encountered several challenges in manually transcribing visual information on slides into captions. Several participants said they must have equations in captions, but they were not sure if their edits were correct or could cause more confusion. Two participants found their previous experience on transcription helpful for the caption editing activity.

Motivation and outcome for editing captions It is not surprising that most participants thought "extra credit" was the main motivation to participate in the caption editing activity. Besides extra credit, participants enjoyed the process of "making things perfect", and meet their habits of "being perfectionism". They felt satisfied upon editing as they improved the validity of caption quality. Three participants(P9, P10, P16) considered their effort to be altruistic behavior that helps peers. They also found their work sustainable for future learners to motivate them to make edits (P9, P10, P17). Participants also reported that closely reading and listening while editing equations and abbreviations helped them prepare for exams(P1–P5, P9). However, some thought the editing task to be discouraging because other learners would not directly recognize their efforts. For example, P8 explained "I don't know who is going to use this caption in the future, you know I don't know who I am helping, the future users won't know my name as well, then why should I do it so carefully?"

Summary: We first provided different types of edits made by examining editing log. We pinned down the decision-making process on making edits through interview data—identifying errors, editing errors, and verifying edits to improve caption editing accuracy and efficiency. We enumerated the editing strategies that were used, identified factors that impact such process—e.g., general usage of the caption, prior domain knowledge, familiarity with transcription task. Our findings showed that *discipline-specific edits* take up 35% of total edits, they were easier to notice, have a major impact on understanding class content, and learners were confident with their editing except *equation edits* . *General edits* formed 65% of total edits, were harder to notice, and learners were less confident with their editing. Learners' domain knowledge and general use of captions also have a great impact on caption editing behavior. Reading slides and re-listening to video clips were two widely used strategies improve accuracy. Guessing and prioritizing edit type was used to improve efficiency.

Collaborative learners' caption editing behavior (RQ2)

As we introduced in the Method section, *editor one* and *editor two* were the two roles participants signed-up for to participate in caption editing activity. By investigating system log, we found five participants edited captions while they were taking online class using current system as *volunteers* . To answer RQ2: How do learners collaborate with other

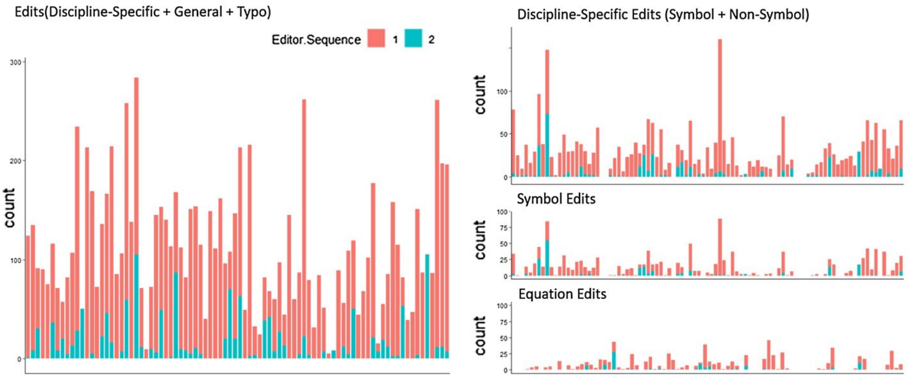


Fig. 4 Numbers of different types of edits made from 89 Videos

learners in crowd-sourced caption editing?, we first analyze on how *editor one* , *editor two* , and *volunteers* collaborate to edit captions using log analysis and interview analysis. Then, we report the editing strategies used in collaborative editing through interview analysis.

Shared responsibility between editor one, editor two and volunteers: log analysis

Below, we report on how *editor one* and *editor two* share responsibility in caption editing by investigating their editing counts, editing type and editing error in editing log. We triangulate our findings by interview questions regarding collaborative editing experience (e.g. "As editor two, what did you think of quality of editor one's work?", "Did your editing standard change while being editor one vs. editor two?")

For this section, we removed edits where participants' actual edited videos differed from the ones they originally signed up for since this could indicate they were confused about whether they were *editor one* or *editor two* for those videos. After data cleaning, we have 8623 edits (91.0% of total edits) made by 54 participants from 89 videos (95.7% of total videos). Every participant signed up for editing two videos as *editor one* and another two as *editor two* , except eight participants who signed-up late and could only sign-up for any remaining available tasks at that time. In Fig. 4, we show *editor one* and *editor two* edits (total edits, *discipline-specific edits* , *symbol edits* , and *equation edits*) for all videos.

Editor one and editor two *Editor one* in total made 7467 edits (Mean= 80.2, SD = 62.5, per video) , while *editor two* in total made 1021 edits (Mean= 11.0, SD = 19.9, per video). *Editor one* made significantly more edits than *editor two* per video (W = 6247.5, p < .001). Same results for *discipline-specific edits* (W = 7130.5, p < .001), *general edits* (W = 7181, p < 0.001), *symbol edits* (W = 6846.5, p < .001), *non-symbol edits* (W = 7176, p < .001). Our analysis also shows *editor one* made significantly more *typos* than *editor two* (W = 6247.5, p < .001). This result alone is not that surprising given that there were likely to be less errors after prior learners made edits. Interview results below reveals *editor two's* explanations for such behavior.

Editor one and *editor two* don't show significant difference in *discipline-specific edits* vs. *general edits* or *symbol edits* vs. *non-symbol edits* . But, within *symbol edits* (*abbreviation edits* , *equation edits* and *single symbol edits*) , *editor one* and *editor two* made significantly different kinds of edits according to Chi-Square testing (X-Square = 7.13, p = .03). Post-hoc results show that *editor two* was more likely to make *single symbol edits* than *editor one* (Residuals = 2.26, p = .05). *Editor one* made 249 *single symbol edits* (24.1%) compared to 61 *single symbol edits* (33.0%) made by *editor two* . Within the 61 *single symbol edits* , a large percentage were editing numbers, for example "*.9*" to "*9*", "*V3*" to "*view three*", "*tool*" to "*2*". Numbers edits were one of the special and challenging cases which required close attention paid to context.

Intrinsically motivated volunteers (did not receive extra-credit for making edits) There were five participants that made 164 edits for in total two videos without signing up as either *editor one* or *editor two* . One of the three Volunteers who participated in the interview said he wasn't aware of the caption editing activity and was purely motivated by interest and enjoyed the "hunting process of making edits" (P17). The participant made six edits for two videos, four General and two Discipline-Specific edits, for example "*water*" to "*world*", and "*engine*" to "*engines*".

It is unsurprising to see that *volunteers* (Mean= 82. SD = 109.1) made comparatively similar amount of total edits as *editor one* (Mean = 80.2) , and more than *editor two* (Mean = 11.0) (W = 1130225, p < .001). Among five *volunteers* , only one *volunteer* edited prior to *editor one* and four *volunteers* edited after *editor one* . Therefore, *volunteers* were not likely to impact *editor one* and *editor two* caption editing task by greatly reducing before-edit error counts.

Volunteers made significantly different types of *symbol edits* than non-volunteers (learners with committed editing tasks) (chi-squared = 50.2, df = 2, p < .001). Volunteers made more *abbreviation edits* that non-volunteers (Residual = 6.8, p < .001). Among all volunteers, three participated in the interview and agreed that they tended to revise errors that were "easy to tell and provide the most accurate edit". P17 :"...error such as *PLA* to *PLSA*, which is really impacting the understanding, a lot of times it was mistaken as several simple words that were obviously not supposed to appear together. "

Collaborative editing strategies

Interviews allowed us to understand better the different strategies used in and factors impacting the collaborative caption editing task.

Firstly, a majority of *editor two* trusted *editor one* 's work and reported they found captions edited by *editor one* is of acceptable accuracy for accessing learning contents. Thus, they sped up videos (1.5 or 2 times original speed) and rarely paused to re-listen. Within this process, they closely compared visual information on slides and captions for better consistency on word formatting, capitalization and numbers in *discipline-specific edits* words. At the same time, they admitted they perceived themselves being "less responsible" as *editor two* compared to *editor one* .

Secondly, we found that learners denoted their unconfidence via a special notation, "*[INAUDIBLE].*" Four learners utilized this strategy. While we were not sure if the learners used the "*[INAUDIBLE]*" notation as an approach to request help from *editor two* , we found two examples in which the second learner helped complete the part denoted as "*[INAUDIBLE]*" by the first editor.

Participant P2, a *volunteer* and a native speaker who replaced an "*[INAUDIBLE]*" notation, commented that "seeing [INAUDIBLE] in others' caption could encourage other learners to make edits." The participant was aware that most learners were non-native speakers, and as a native speaker, P2 was more confident in making such edits. Even though a few participants agreed that using special notations such as "*[INAUDIBLE]*" helped communicate confusion and draw the attention of *editor two* , three interviewees (P4,P7,P9) thought that using notations such as "*[INAUDIBLE]*" may create confusion for general learners.

Thirdly, even though participants reported applying the same editing standard regardless of their editor roles, most of them said that they needed more confidence for fixing the same error as *editor two* compared to *editor one* . One reason for the need of additional confidence, as put succinctly by P7, is that as a second editor, "I don't want to step on someone else's toes." Another concern of second editors was adding personal styles that were different from those of the first editors, which could cause inconsistency and confusion.

Summary: Our findings show that *editor one* , *editor two* and *volunteers* contributed to caption editing in a complementary manner. *editor one* made more edits and spent more time. *Editor two* made fewer edits, made significantly different types of *discipline-specific edits* than *editor one* . *editor two* were more likely to make *single symbol edits* , mainly edits on numbers. *Editor two* considered efficiency to be more important than *editor one* . *Editor two* often over-trusted the work of *editor one* and had lower self-efficacy than *editor one* . Learners used special notations, e.g., "*[INAUDIBLE]*", to communicate uncertainty in their edits, and such notations resulted in further updates from subsequent editors.

ML-based solutions for better system support (RQ3)

To answer RQ3: How can the system better support learners to conduct caption edits for educational videos?, we investigate two perspectives. First, we asked interviewees for their suggestions of what additional support they would like during the caption editing task; the results are presented in "Interviewee suggestions to improve caption editing experience". Second, based on our findings from RQ1 and participants' suggestions, we developed a machine learning model to help with error verification and evaluated its feasibility.

Interviewee suggestions to improve caption editing experience

In the interview, we asked participants for their suggestions on how to improve the crowd-sourced caption editing task, including the interface, the process, and the policies for the task. In this section, we zoom-in interviewees' challenges that lacks direct technology solutions provided by participants.

Challenges that participants has provided technology solutions on: three interviewees (P9, P17, P20) proposed that incorporating grammar checks could help them identify potential edits more easily. P16 suggested to improve the user interface to minimize the number of clicks required for each edit. Three interviewees explicitly mentioned that providing editing guidelines would be very helpful, ranging from defining the balance between transcribing everything (including speaker errors) versus optimizing the readability of the caption, to specifying transcription styles such as when to capitalize a word (P1, P2, P16). In the course being studied, learners were asked to edit the captions before the final exam. While some learners found the task helpful for preparing for the final exam as discussed in "Individual editing strategies", seven participants preferred editing the captions at their own time while watching the videos. They thought that editing captions before an exam felt like extra labor, while editing when they proactively watch a video any time in a semester

Table 4 Performance of classifier-verify on test set

Classifier	Method	Acc.(%)	Discipline-Specific			General		
			P	R	F1	P	R	F1
Baselines	Majority	67.6	0.0	0.0	0.0	0.68	1.0	0.81
	Random	52.1	0.35	0.53	0.42	0.68	0.48	0.56
CV	RF	85.4	0.92	0.6	0.73	0.83	0.99	0.90
	LR	**93.8**	**0.92**	**0.89**	**0.90**	**0.94**	**0.95**	**0.94**

Bold numbers indicate the best performance

would feel more organic and hence more motivating. In addition to changing the timing of the task, three participants (P5, P7, P8) thought that visualizing learners' editing contributions could motivate editing behavior as well.

New learners lacked the ability to identify and verify discipline specific errors: A major factor that impacted learners' ability to identify errors was their prior domain knowledge (4.1.2). Prior domain knowledge was also a factor that impacted participants' confidence in their edits (4.1.2). P11 suggested that learners could benefit from an overview of the main concepts mentioned in a particular video before the editing task, e.g., the explanation of a keyword. Along the same line, P3 suggested that they would be more confident with the caption quality if their own edits could be checked by learners who have good domain knowledge. P11's suggestion of providing an overview of concepts mentioned in a video could also build confidence in a learner during editing. In short, although new learners are able to edit the error, it could be challenging to identify and verify the errors. Such findings indicate the need to differentiate error editing types and assign to learners with different domain knowledge.

Tech assistance was required for repetitive discipline specific error editing: As shown in our findings of caption error editing (4.1.2), interviewees dealt with a wide array of edit types. Some of them expressed the needs for more support when editing discipline specific errors. Three interviewees (P1, P5, P15) asked for features to help avoid repetitive editing on the same errors, such as automatic correction. Two interviewees (P9, P11) wanted more support for *equation edits* , such as enabling Greek alphabet and special symbols. Individual interviewees requested improvement for less common but more complicated edits. P20 thought auto-completion suggestions could be helpful in reducing editing efforts. Such findings indicate the need to differentiate error editing types and provide different technical solutions accordingly.

Machine learning model for efficient edit verification

Learners' suggestions in previous section and findings in RQ1 indicate that there is a need to reduce and optimize their effort during manual editing.

Thus, we trained machine learning models leveraging our coded log data described in "Machine learning model for supporting caption editing". Our goal is to build a classifier to help optimize efforts during verification by prioritizing either *discipline-specific* or *general edits* based on their preferences. We refer to this classifier as Classifier-Verify (CV). The task of Classifier-Verify model is posed as a binary classification problem, where given a word w_{aft}

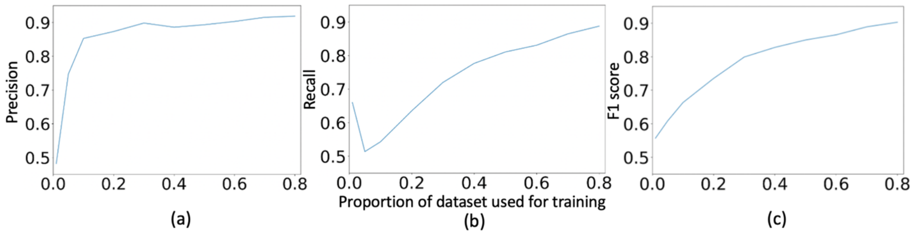


Fig. 5 Variance of **a** Precision, **b** Recall and (**c**) F1 scores with training set size for the Discipline-Specific class on the test set using CI

and c_{aft}, the classifier predicts whether w_{aft} is a *discipline-specific edits* or *general edits* . In this section, we discuss the feasibility of such models.

Table 4 shows the results (overall accuracy, Precision (P), Recall (R), and F1-measure (F1) for both *discipline-specific* and *general* class using CV on the test set. As can be seen from the Majority baseline accuracy, *general* class constitutes about 68% of the test set. Logistic Regression (LR) achieves the best overall performance for both *discipline-specific* and *general* classes. The high performance is likely due to many repeated errors and edits as discussed before, specifically w_{aft} words (86%) in the test set were also present with the training set. This suggests an opportunity to leverage ML models for this task.

We now investigate how much training data is required to achieve a comparable performance as obtained by using 80% of the dataset. Figure 5 shows the variance of test set performance on the Discipline-Specific class of classifier *CV* trained on the earliest (based on timestamps of edits) *x*% of the word-level edits, where *x* ranges from 1 to 80. Performance on the *general* class follows similar patterns only with higher scores. Here, we used the best performing method, Logistic Regression with hyperparameters tuned on the whole training set. We can see that performance rises sharply and then almost plateaus. Interestingly, the precision of a model trained on about 10% (1k edits or equivalently edits made on about 6 – 7 lecture videos) of the total edits, is already quite close in performance (within 10%) to that trained on the whole training set (80% of the dataset). As expected, with more data, especially the recall and F1 get a further boost.

Summary: Participants suggested several ways to better support both individual and collaborative caption editing during the Find (Identify), Fix and Verify stages. Inspired by those suggestions and findings in RQ1, we designed a classifier. Our model evaluation findings show that it is feasible to develop machine learning models to optimize and reduce effort during error verification. The CV model to prioritize errors based on their types, during the Verify stage, achieves a high performance and could be tested for use in real-world systems. Using only about 1k labelled edits helps achieve a high-precision model and adding more training data could further boost coverage of detected *discipline-specific* and *general edits* . We note that the models trained are more of a proof of concept and need further tuning and testing before they can be deployed in a real world scenario.

Discussion

Design implication for online lecture videos caption editing

In this section, we propose design implications for improving captions in educational videos based on the findings of RQ1, RQ2 and RQ3, which suggest a suite of captioning needs that are unique to caption-editing for educational videos. For example, learners need to correct repetitive errors. So, a batch find and replace feature to quickly fix all instances of a repetitive error (e.g.,"*cedar*" → "*theta*") could be useful to help save future learners' time and effort compared to editing each instance of the error individually. There are multiple such unique challenges for either ASR or human to correctly transcribe educational videos. Accordingly, we suggest opportunities for improvement, particularly for developing and improving machine learning algorithms for automated and machine-assisted caption generation, and sustaining collaborative caption editing by learners.

Machine learning models for improving captions

For improving captions for educational videos, we identify new opportunities for developing better machine learning algorithms. Using end-user input to improve the performance of machine learning models has been shown to help in building intelligent systems (Amershi et al., 2014). Our findings show several interesting editing behaviors on algorithmically generated captions that could be leveraged to improve such algorithms (in our case, Azure). One important finding is that punctuation-related learner edits are important for learning, e.g. due to incorrect caption segmentation. Future algorithms could also learn from sufficient learners' edits to improve errors and caption segmentation performance which is challenging when purely decoding audio files (Alvarez et al., 2017).

Secondly, in addition to using the editing log data to directly improve ASR algorithms, it could also be used to further assist learners in editing. Such human-in-the-loop models provide more control to learners, thus ensuring higher accuracy, while reducing and optimizing their editing efforts. In Section "Machine learning model for efficient edit verification", we noticed that that a high-accuracy classifier (Classifier-Verify) can be developed for classifying words inserted (or substituted) by learners as *discipline-specific* or *general* . Such a classifier could be used for providing the aforementioned control over edit-type prioritization. We note that the edited words need to be manually labelled as *discipline-specific* or *general* to create a training set. However, in Section "Machine learning model for efficient edit verification", we observed that it might be possible to deploy a high-precision classifier trained on a reasonably small sized edit log dataset early during the course. Adding more training data could further help boost the recall (i.e. increase coverage of identified *discipline-specific* or *general* edits).

Sustaining collaborative editing by learners

We also have several points to reflect on how to sustain collaborative caption-editing behavior in learners and encourage them to make more edits.

Firstly, we should make the editing task better serve the purpose "to learn". For example, it should not overload and increase unnecessary cognitive load. Future work could also explore designing separate learning and editing modes on the system to avoid any distractions due to editing during learning.

Secondly, our findings under RQ-3 explicitly suggested such editing task could be "gamified" and "engaging", e.g., show animations to create error "hunting" environment. Previous research also found that some gamification elements, such as badges and leaderboards, can lead crowdworkers to do more work than they are paid for (Lichtenberg et al., 2020).

Thirdly, under RQ2-3, our findings suggest that participants are motivated when 'healthily competing with others' and 'are recognized by others'. Further research could investigate how to utilize counseling mechanisms to increase editing performance, such as 'Strength-Based Counseling Mode' motivating individuals to embrace strengths they may have when encountering adversity in the pursuit of higher goals (Smith, 2006).

Fourthly, some learners who knew that they were editing at later sequence felt less responsible and struggled to be more patient in editing task (RQ2). Future learner-sourcing systems without task division should design mechanisms to moderate such effects and facilitate effective coordination and teamwork.

Fifthly, our findings show that learners without sufficient knowledge could be confused and less confident in making edits. Motivated by the use of notations like "*[INAUDIBLE]*" that we observed in our study, one possibility could be to add a flag feature to future systems, where learners (e.g. *editor one*) could flag a segment or word that they are confused with to communicate their uncertainty to subsequent learners editing it. Moreover, pre-screening methods have been utilized to select appropriate crowd workers for a given task (Gadiraju et al., 2019). Besides self-assessment and behavior modeling, further editing systems for online video lecture caption editing could also use algorithms, e.g. automatic question generation (AQG) techniques (Kurdi et al., 2020) to generate lecture-related question to pre-screen suitable learners.

Limitations and future work

This study contains several limitations that can benefit from future research. First, the study was conducted with only one course in the Computer Science department on the topic of text mining, taught by an instructor who is a non-native English speaker. All participants used ClassTranscribe for captioning, which might have affected captioning behaviors and practices of learners. The caption error taxonomy we presented in this study were specific to captions in the topic of text mining that were transcribed from audio spoken by a non-native speaker. Future studies on courses across different topics and disciplines and taught by instructors with different accents are needed to further expand the taxonomy of caption errors in educational videos.

Second, our interview sample was biased towards non-native English speaker and didn't include any learner with chronic physical or mental health conditions that would prevent them from understanding the lecture video content. Both of these characteristics impact learners' ability to edit captions and how much they valued accurate captions. Future work can further investigate such impacts by hearing more from native speakers and learners with special needs.

Lastly, we developed the machine learning models with the goal of evaluating the feasibility of the machine learning-based solutions to support crowdsourced caption editing. Therefore, the models are more of a proof of concept and need further tuning and testing before they can be deployed in a real world scenario. Future work can improve the performance of our models and test their effectiveness in various use cases.

Conclusion

In this paper, we present our study on learner-sourcing to edit educational video captions. Our study deployed a system for editing lecture video captions in a large enrollment (N=387) text mining course where 58 learners participated in editing captions of 89 lecture videos. Each lecture video was edited by two learners sequentially. Eighteen of those learners participated in follow-up interviews. From analysing system edit logs and qualitative analyses, we found that there is a taxonomy of errors in educational video captions. Moreover, participants used varied individual and collaborative strategies while editing the different types of errors. Inspired by the findings and learners' suggestions for better system support, we evaluated the feasibility of a proof-of-concept machine learning model to assist learners identify and prioritize *discipline-specific* and *general* errors during the Verify stage. We also discussed the practical implications and system design suggestions based on our findings.

Funding This material is based upon work supported by the National Science Foundation under Grant Nos. 2119589 and 1801652. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This project was also made possible in part by the Institute of Museum and Library Services LG-70-18-0088-18. The views, findings, conclusions or recommendations expressed in this article do not necessarily represent those of the Institute of Museum and Library Services. The research reported here was also supported by a Microsoft Corporation gift to the University of Illinois as part of the 2019 Lighthouse Accessibility Microsoft-Illinois partnership.

Declarations

Ethical approval Approval for this study was granted by the IRB at UIUC (IRB#: 21526)

Consent to participate Informed consent was obtained from all individual participants included in the study.

References

Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. *Mining text data* (pp. 163–222). Springer.

Alvarez, A., Martínez-Hinarejos, C.-D., Arzelus, H., Balenciaga, M., & del Pozo, A. (2017). Improving the automatic segmentation of subtitles through conditional random field. *Speech Communication, 88,* 83–95.

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. *Ai Magazine, 35*(4), 105–120.

Amos, J. R. , Zhang, Z. , Angrave, L. , Liu, H., & Shen, Y. (2021). A udl-based large-scale study on the needs of students with disabilities in engineering courses. In *2021 asee virtual annual conference content access.*

Angrave, L., Jensen, K., Zhang, Z., Mahipal, C., Mussulman, D., Schmitz, C. D., & Kooper (2020a). Improving student accessibility, equity, course performance, and lab skills: How introduction of classtranscribe is changing engineering education at the university of illinois. In *Asee annual conference & exposition*

Angrave, L., Zhang, Z., Henricks, G., & Mahipal, C., (2020b). Who benefits? Positive learner outcomes from behavioral analytics of online lecture video viewing using classtranscribe. In Proceedings of the 51st acm technical symposium on computer science education (p. 1193-1199). Association for Computing Machinery. https://doi.org/10.1145/3328778.3366953

Bernstein, M. S. , Little, G. , Miller, R. C. , Hartmann, B. , Ackerman, M. S. , Karger, D. R., & Panovich, K. (2010). Soylent: A word processor with a crowd inside. In *Proceedings of the 23nd annual acm symposium on user interface software and technology* (pp. 313–322)

Bozzon, A., Mauri, A., & Brambilla, M. (2012). A model-driven approach for crowdsourcing search. In *Crowdsearch*. workshop at WWW 2012, Lyon, France (pp. 31–35)

Clossen, A. S. (2014). Beyond the letter of the law: Accessibility, universal design, and human-centered design in video tutorials. *Pennsylvania Libraries: Research & Practice, 2*(1), 27–37.

Cross, A., Bayyapunedi, M., Ravindran, D., Cutrell, E., & Thies, W. (2014). Vidwiki: Enabling the crowd to improve the legibility of online educational videos. In *Proceedings of the 17th acm conference on computer supported cooperative work & social computing* (pp. 1167–1175)

Cross, J. S., Keerativoranan, N., Carlon, M. K. J., Tan, Y. H., Rakhimberdina, Z., & Mori, H. (2019). Improving mooc quality using learning analytics and tools. In *2019 ieee learning with moocs (lwmoocs)* (pp. 174–179)

Culbertson, G., Shen, S., Andersen, E., & Jung, M. (2017). Have your cake and eat it too: Foreign language learning with a crowdsourced video captioning system. In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing* (pp. 286–296)

Deshpande, R., Tuna, T., Subhlok, J., & Barker, L. (2014). A crowdsourcing caption editor for educational videos. In *2014 ieee frontiers in education conference (fie) proceedings* (pp. 1–8)

Dye, M., Nemer, D., Kumar, N., & Bruckman, A. S. (2019). If it rains, ask grandma to disconnect the nano: Maintenance & care in havana's streetnet. Proceedings of the ACM on human-computer interaction, *3*(CSCW), 1–27.

Dym, B., Brubaker, J. R., Fiesler, C., & Semaan, B. (2019). coming out okay community narratives for lgbtq identity recovery work. *Proceedings of the ACM on human-computer interaction, 3*(CSCW), 1–28.

Errattahi, R., El Hannani, A., & Ouahmane, H. (2018). Automatic speech recognition errors detection and correction: A review. *Procedia Computer Science, 128,* 32–37.

Filippidou, F., & Moussiades, L. (2020). A benchmarking of ibm, google and wit automatic speech recognition systems. In *Ifip international conference on artificial intelligence applications and innovations* (pp. 73–82)

Gadiraju, U., Demartini, G., Kawase, R., & Dietze, S. (2019). Crowd anatomy beyond the good and bad: Behavioral traces for crowd worker modeling and pre-selection. *Computer Supported Cooperative Work (CSCW), 28*(5), 815–841.

Hamilton, L., & Corbett-Whittier, C. (2012). *Using case study in education research*. Sage.

Hrinchuk, O., Popova, M., & Ginsburg, B. (2020). Correction of automatic speech recognition with transformer sequence-to-sequence model. In *Icassp 2020-2020 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 7074–7078)

Huang, Y., Huang, Y., Xue, N., & Bigham, J. P. (2017). Leveraging complementary contributions of different workers for efficient crowdsourcing of video captions. In *Proceedings of the 2017 chi conference on human factors in computing systems* (pp. 4617–4626)

Klein, R. (2021). U.s. laws for video accessibility: Ada, section 508, cvaa, and fcc mandates. Retrieved December 12, 2021, from https://www.3playmedia.com/blog/us-laws-video-accessibility/

Klimt, B., & Yang, Y. (2004). Introducing the enron corpus. In *Ceas* 2004 - First Conference on Email and Anti-Spam, July 30–31, 2004, Mountain View, CA

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering, 30*(1), 25–36.

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education, 30*(1), 121–204.

Lewis, E. (2021). Captioning and transcription for stem content. Retrieved December 12, 2021, from https://www.3playmedia.com/blog/captioning-and-transcription-for-stem-content/

Lichtenberg, S., Lembcke, T., Brenig, M., Brendel, A., & Trang, S. (2020). Can gamification lead to increase paid crowdworkers output? 15. *Internationale Tagung Wirtschaftsinformatik*

Mahipal, C., Angrave, L., Xie, Y., Chatterjee, B., Wang, H., & Qian, Z. (2019). what did i just miss?! presenting classtranscribe, an automated live-captioning and text-searchable lecture video system, and related pedagogical best practices. In *2019 asee annual conference & exposition*. Tampa, Florida: ASEE Conferences. https://peer.asee.org/31926

McCarron, L. (2021). Creating accessible videos: Captions and transcripts. *Communications of the Association for Information Systems, 48*(1), 19.

Merriam, S.B. (1985). The case study in educational research: A review of selected literature. *The Journal of Educational Thought (JET)/Revue de la Pensée Educative),* 204–217.

Parton, B. (2016). Video captions for online courses: Do youtube's auto-generated captions meet deaf students' needs? *Journal of Open, Flexible, and Distance Learning, 20*(1), 8–18.

Ranchal, R., Taber-Doughty, T., Guo, Y., Bain, K., Martin, H., Robinson, J. P., & Duerstock, B. S. (2013). Using speech recognition for real-time captioning and lecture transcription in the classroom. *IEEE Transactions on Learning Technologies, 6*(4), 299–311.

Ren, J. C. , Hasegawa-Johnson, M., & Angrave, L. (2015). Classtranscribe: a new tool with new educational opportunities for student crowdsourced college lecture transcription. In *Slate* (pp. 179–180)

Shiver, B. N., & Wolfe, R. J. (2015). Evaluating alternatives for better deaf accessibility to selected web-based multimedia. In *Proceedings of the 17th international acm sigaccess conference on computers & accessibility* (pp. 231–238)

Smith, E. J. (2006). The strength-based counseling model. *The Counseling Psychologist, 3*(4), 113–179.

Tellis, W. (1997). Introduction to case study. *The Qualitative Report, 269*

Valor Miró, J. D., Spencer, R. N., González, Pérez., de Martos, A., Garcés Díaz-Munío, G., Turró, C., et al. (2014). Evaluating intelligent interfaces for post-editing automatic transcriptions of online video lectures. *Open Learning: The Journal of Open, Distance and e-Learning, 29*(1), 72–85.

Wald, M. (2013). Concurrent collaborative captioning. In *Proceedings of the International Conference on Software Engineering Research and Practice SERP'13*. CSREA Press.

Wilcoxon, F. (1992). Individual comparisons by ranking methods. *Breakthroughs in statistics* (pp. 196–202). Springer.

Zhang, Z. (2021). Attitudes, behaviors, and learning outcomes from using classtranscribe, a udl-featured video-based online learning platform with learnersourced text-searchable captions (Unpublished doctoral dissertation)

Zhang, Z., Bhavya, B., Angrave, L., Sui, R., Kooper, R., Mahipal, C., & Huang, Y. (2021). How students search video captions to learn: An analysis of search terms and behavioral timing data. In *2021 asee virtual annual conference content access*

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Bhavya Bhavya is a second year Ph.D. student in the Department of Computer Science at the University of Illinois at Urbana-Champaign. She is broadly interested in human-centered machine learning, particularly for developing intelligent educational systems.

Si Chen is a third year Ph.D. student in the School of Information Sciences at the University of Illinois at Urbana-Champaign. She is interested in human-computer interaction, particularly for designing inclusive and intelligent educational systems.

Zhilin Zhang is a CS student at UIUC. His research interests are broadly in Human-Computer Interaction (HCI). He studies HCI to better understand and improve human interactions with AI, VR/AR, and other complex algorithmic systems. He designs and builds computing systems to positively support users' online behaviors and interactions in a scalable and accessible way.

Wenting Li is a 4th year Ph.D. student in the Computer Science Department at UIUC, advised by Dr. Hari Sundaram and Dr. Karrie Karahalios. Her research interests broadly lie in the intersection of human-computer interaction (HCI), education technology, and artificial intelligence (AI). She is excited in exploring the role of AI in providing quality, equitable, and accessible education at scale.

Chengxiang Zhai is a Donald Biggar Willett Professor in Engineering in the Department of Computer Science at the University of Illinois at Urbana-Champaign. His general interests are in developing novel Intelligent Information Systems (e.g., intelligent search engines, recommender systems, text analysis engines, and intelligent task assistants) to help people manage and exploit large amounts of data (i.e., "big data"), especially text data.

Lawrence Angrave is an award winning Fellow and Teaching Professor at the department of computer science at the University of Illinois at Urbana-Champaign (UIUC). His interests include (but are not limited to) joyful teaching, empirically-sound educational research, campus and online courses, computer science, unlocking the potential of underrepresented minorities, improving accessibility and creating novel methods that encourage new learning opportunities and foster vibrant learning communities.

Yun Huang is faculty in the School of Information Sciences at the University of Illinois at Urbana-Champaign. Her expertise is in the area of social computing, human-computer interaction, Internet of Things, and human-AI interaction. In her work, she designs, implements and evaluates social computing systems that can engage community members to co-create new services for better community wellbeing.