Using Vocabularies to Collaboratively Create Better Plans for Writing Tasks

Harmanpreet Kaur
University of Michigan
harmank@umich.edu

Walter S. Lasecki
University of Michigan
wlasecki@umich.edu

Alex C. Williams
University of Waterloo
alex.williams@uwaterloo.ca

Shamsi Iqbal
Microsoft Research
shamsi@microsoft.com

Anne Loomis Thompson
Microsoft Research
annelo@microsoft.com

Jaime Teevan
Microsoft Research
teevan@microsoft.com

Abstract

Having a step-by-step list of instructions for completing a task—a plan—enables people to make progress on challenging tasks, but making plans for tasks is a tedious job. Asking crowdworkers to make plans for others' tasks only works for independent (context-free) tasks, and asking people who have context (e.g., friends or collaborators) has social costs and quality concerns. Our goal is to reduce the costs and improve quality of planning by people who have context in the context-rich domain of writing. We introduce a vocabulary (a finite set of functions pertaining to writing tasks) to aid the planning process. We develop a writing vocabulary by analyzing 264 comments, and compare plans created using this vocabulary to those created without any aid, in a study with 768 comments ($N = 145$). We show that using a vocabulary reduces the planning time and effort compared to unstructured planning, and opens the door for automation and task sharing for complex tasks.

Author Keywords

Action plans; task decomposition; crowdsourcing

ACM Classification Keywords

H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous

CHI'18 Extended Abstracts, April 21–26, 2018, Montreal, QC, Canada
ACM 978-1-4503-5621-3/18/04.
https://doi.org/10.1145/3170427.3188640

Background and Related Work

Having a plan — a step-by-step list of instructions for how to implement a task — helps people get started on tasks and complete them faster [2, 5]. While making plans for one's tasks is considered a helpful form of contemplation, people prefer to have automatically created plans at their disposal over making them. Prior work shows that people complete more tasks when they have access to personalized plans for their tasks that are created by others [4]. We also see evidence of this in the rise of industry applications that provide planning services: RunKeeper plans people's workouts based on a health goal, Cook Smarts helps plan their meals, and Mint helps plan finances.

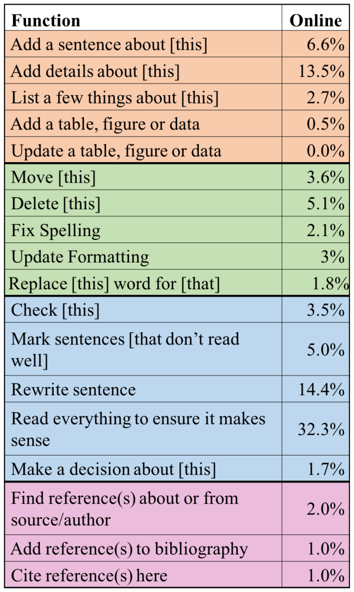
While having personalized, pre-generated plans is beneficial and preferred, it is hard to outsource plan creation for all kinds of tasks. This is especially true for context-embedded tasks – tasks that require additional contextual information, such as comments/to-dos in a written document or codebase, health goals, etc. Prior work has employed crowd workers to make plans for independent tasks (e.g., planning a trip from an airport to a hotel) because these can be planned without additional information. However, crowd workers or, more generally, people who do not have the required context, cannot plan for context-embedded tasks [4]. Even people who already have the required context (e.g., collaborators, friends) are often not a feasible option. People worry about asking for help with planning because of social costs: they do not want to share all personal information with friends, or ask collaborators to spend time and effort to do the majority of work for their task [1]. Additionally, people with context can be biased due to this inside information, and not include all the basic steps necessary to accomplish a task. Applying fully-automated approached also does not work for context-embedded task planning due to lack of natural language understanding.

In this paper, we explore a mechanism for outsourcing context-embedded task planning to people with some context (e.g, collaborators), while mitigating the costs and quality concerns mentioned above. We develop a vocabulary — a set of basic functions that make up larger tasks in a domain — and study the effects of using it for planning. This vocabulary provides a language for breaking down a larger task into a set of actionable steps. Bootstrapping the cognitive process of planning in this way can reduce the time and effort costs on collaborators, and having a set of basic functions to choose from could reduce instances where all basic steps are not listed. We focus on the domain of writing, exploring how comments left in a document can be transformed into a plan. We choose comments in a document as a proxy for writing tasks since comments provide a set of interdependent, context-rich tasks that need to be accomplished to improve the writing.

Our project has two phases. First, we qualitatively create a vocabulary of 18 functions for accomplishing writing tasks, by using 264 comments left on Wikipedia articles and academic papers. We then compare how people with context make vocabulary-based plans vs. unstructured plans (wherein plans are made with no aid), via a study with 145 Mechanical Turk workers creating plans for 768 comments on Wikipedia articles. Our results indicate that people spend less time and effort making vocabulary-based plans compared to unstructured ones, and make vocabulary-plans with more broken down, basic steps.

Measures of Success

Our approach is to generate a list of functions a priori – this list can then be used by people to come up with the steps of a plan. We call this list of functions a vocabulary. We compare our vocabulary-based planning approach to an unstructured planning process, wherein plans are made



Function	Online
Add a sentence about [this]	6.6%
Add details about [this]	13.5%
List a few things about [this]	2.7%
Add a table, figure or data	0.5%
Update a table, figure or data	0.0%
Move [this]	3.6%
Delete [this]	5.1%
Fix Spelling	2.1%
Update Formatting	3%
Replace [this] word for [that]	1.8%
Check [this]	3.5%
Mark sentences [that don't read well]	5.0%
Rewrite sentence	14.4%
Read everything to ensure it makes sense	32.3%
Make a decision about [this]	1.7%
Find reference(s) about or from source/author	2.0%
Add reference(s) to bibliography	1.0%
Cite reference(s) here	1.0%

Figure 1: Our vocabulary of writing primitives and the percentage of times each primitive is used in our study. The primitives are grouped by purpose: adding content (top), surface-level issues (second), editing content (third), and references (bottom).

with no oversight or defined process, using five measures: *time, effort, granularity, atomicity,* and *sourcing potential.*

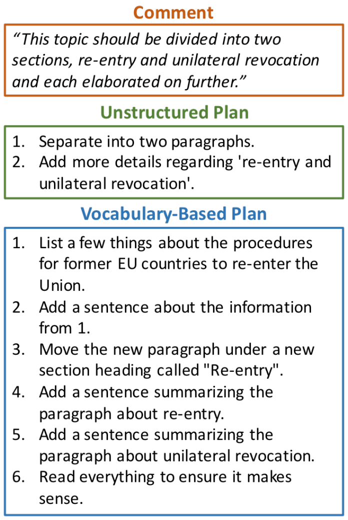
We measure the time it takes for people to make plans and compare it for conditions with- and without-vocabulary. To measure effort, we use the NASA-Task Load Index (TLX) questionnaire. Ideally, plans are a list of microtasks that accomplish a macrotask – one way to measure plan quality is by checking how well the original task is broken down. We use three proxies for this: (i) the number of steps generated per plan (granularity); (ii) the number of steps of the plan that are mechanical, i.e., can be done with little cognitive effort or need for context (atomicity); and (iii) the number of steps that can be crowdsourced or selfsourced, i.e., can be done in micromoments (sourcing potential).

Phase 1: Vocabulary Creation

Our first step was to create a vocabulary that consists of some basic functions for writing tasks. We followed an inductive, data-driven qualitative coding process to create this vocabulary. We used two types of articles – academic and general information entries – to create the vocabulary. We collected 10 documents written for academic audiences: five summer project descriptions written by interns at a large technology company, and five nearly complete drafts of papers being submitted to various HCI conferences. For the five intern project descriptions, we ask project mentors to provide feedback in the form of comments on these documents, whereas for the nearly complete drafts, this feedback was already included in the drafts collected.

We also included general writing instances in the form of Wikipedia articles. We picked the top five most popular Wikipedia categories (Popular Culture, Geography, Arts, History, and Current Events), and queried Wikipedia databases to get articles belonging to these categories. For each category, we picked articles graded as "Start" (incomplete articles still in development phase) or "C" (substantial articles, but missing important content and containing irrelevant material) because these are the two longest stages in the lifecycle of an article according to Wikipedia's grading scheme. Additionally, we ensured that the selected articles were at least one page long to get a meaningful amount of text. We were unable to find any articles that overcame the constraints for the Current Events category, giving us eight articles in all. For each article selected, we recruited four Amazon Mechanical Turk workers to provide comments (total 4x8 = 32 workers, pay = $1.50). To ensure a balance among different kinds of writing tasks, we instructed mentors and crowd workers to leave at least two comments for each of the following categories per document (total 10 academic and 8 Wikipedia): (i) mechanics, or surface-level details, such as grammar, spelling, or presentation of repetitive ideas; (ii) organization, or how the content is structured into various sections and paragraphs; and (iii) semantics, or meaning-making, ensuring that the content makes sense, explains the topic, and is not missing details. These categories and their definitions are borrowed from the rhetorical writing categories identified by Greer et al. [3].

We generated the vocabulary by first making step-by-step action plans for all the comments in our dataset (150 academic, 114 general = total 264 comments). Per Zacks et al. [6], we recursively broke down each step of a plan until it was the most primitive function we could identify. We created a list of functions used for each plan, and iterated over it to remove redundant functions and break functions down further. After each iteration, we updated the plans per document in accordance with the new list. We followed this inductive, iterative qualitative process until we had a list of functions based on plans for nine out of 10 academic documents, and seven out of eight Wikipedia articles. We



Comment

"This topic should be divided into two sections, re-entry and unilateral revocation and each elaborated on further."

Unstructured Plan

1. Separate into two paragraphs.
2. Add more details regarding 're-entry and unilateral revocation'.

Vocabulary-Based Plan

1. List a few things about the procedures for former EU countries to re-enter the Union.
2. Add a sentence about the information from 1.
3. Move the new paragraph under a new section heading called "Re-entry".
4. Add a sentence summarizing the paragraph about re-entry.
5. Add a sentence summarizing the paragraph about unilateral revocation.
6. Read everything to ensure it makes sense.

Figure 2: An example comment, and the unstructured and vocabulary-based plans created for it in our study.

used the final documents to conduct an inter-rater reliability test for making plans using our function list (i.e., our vocabulary). The first author and a collaborator made plans for comments in these documents using our vocabulary, and calculated inter-rater reliability on these plans. The plans had substantial agreement with a Cohen's Kappa of 0.72, giving us a final vocabulary of 18 functions (Figure 1).

Phase 2: Plan Creation

We compare plans made using the vocabulary above with unstructured plans in a study using Wikipedia articles with Amazon Mechanical Turk workers. Our study is comprised of three steps: obtaining documents, leaving comments on documents, and making plans for the comments left.

Obtaining documents. We use Wikipedia articles for commenting and planning. We pick the top 16 topical categories and two Start class articles for each category.

Leaving comments. We ask crowdworkers to read and leave comments on the selected Wikipedia articles (pay = $1.50). To get a consistent number of comments throughout the article, we divide each article into two sections. Each section is assigned to two crowdworkers, and each crowdworker is asked to leave six comments (as before, two comments each for issues related to mechanics, organization and semantics). This results in a total of 768 comments generated by 128 crowdworkers (24 comments per article).

Making plans. We ask a different set of crowdworkers to make plans for addressing each comment per article, in a survey format built on SurveyGizmo. We use crowd workers as a proxy for Wikipedia editors since the encyclopedia can be written and edited by anyone. We assign workers to either unstructured planning or vocabulary planning condition. Both conditions have the same steps, modified slightly with the vocabulary. First, we ask crowdworkers to skim the article in $\sim$10 minutes to gain some context about the topic and article. Second, crowdworkers go through a training to make plans. For the vocabulary condition, we train crowdworkers by providing our vocabulary functions and descriptions, and showing them example plans made using the vocabulary. We also ask them to make a sample plan, and show them our plan for the same comment, to validate their understanding of the use of our vocabulary. For the unstructured condition, we have no aid for the crowdworkers, but we show them the same example plans as those in the vocabulary condition. After training, crowdworkers answer the NASA-TLX questionnaire to measure effort. Next, each crowdworker is assigned to one out of two sections per article, and is asked to make plans for the 12 comments left in that section. The vocabulary condition provides our vocabulary as a planning aid whereas the unstructured condition has no aid. Crowdworkers then fill out a NASA-TLX questionnaire about the planning task and answer some open-text questions about the entire process. They are compensated with $6 for the entire study.

Dataset. Each article was planned for by four crowdworkers (two per section – one for unstructured, one for vocabulary-based planning). Crowdworkers were randomly assigned to an article section and a planning condition. Due to a skew in the random assignment process of our survey tool, SurveryGizmo, we generated 840 unstructured plans and 900 vocabulary-based plans. In comparing the two conditions, we ensure that the pairwise comparison is done for each unstructured-vocabulary plan pair obtained in our study, giving us a total of 1014 comparison points.

Results

Figure 2 presents an example from our study: a comment and the unstructured and vocabulary-based plans for it.

Time. People take 120 seconds to make an unstructured

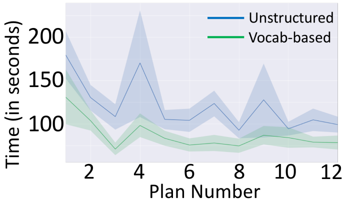


Figure 3: Time taken for planning from Plan 1 to Plan 12. Time spent decreases as planners go from Plan 1 to 12. Vocabulary-based planning takes less time than Unstructured planning at all points.

plan and 87 seconds to make a vocabulary-based plan, on average (s.d. unstructured = 210, vocabulary = 113). The median times are 83 seconds and 57 seconds, respectively. Results of a linear mixed effects model indicate that vocabulary-based planning takes significantly less time compared to unstructured planning ($p < 0.01$). The model uses time as the dependent variable, unstructured vs. vocabulary condition as the independent variable (fixed effects) and participants as the random effects. Participants are random effects since each participant provided 12 datapoints in our dataset – each planned for 12 comments. This implies that our vocabulary benefits planners by reducing the time it takes to make plans. We further compare the amount of time it takes to make each plan per condition to observe learning effects, i.e., whether planning gets easier over time. This is particularly relevant in the vocabulary condition since there is no longitudinal training with the vocabulary—we ask crowdworkers to use it immediately after being introduced to it. We find that: (i) vocabulary planning consistently takes less time than unstructured planning; and (ii) there is a decreasing trend in planning time for the vocabulary condition (Figure 3). Our hypothesis is that using the vocabulary gets easier over time due to a learning effect (see margin for participant quotes supporting this).

Participant Quotes About Vocabulary Use Over Time:

“I think I got the hang of the step-by-step process for addressing comments. It got easier as I went on.” (P7)

“I liked having the guide [vocabulary] to make the instructions from. It took a while, but I got used to it eventually.” (P108)

Effort. We use the NASA-TLX questionnaire data to compare the effort required to train and plan in the unstructured vs. vocabulary conditions. To make these comparisons, we conduct Mann-Whitney’s U tests on each NASA-TLX question for both training and planning tasks. For the training task, people felt significantly more hurried by the pace of the unstructured training compared to the vocabulary training ($p < 0.005$), and significantly more insecure and annoyed during unstructured training ($p < 0.05$). We did not expect this since unstructured training did not involve as many components as the vocabulary training. However, we hypothesize that this outcome is due to differences in the level of engagement in the two trainings: unstructured training simply presented some example plans and asked people to read through these carefully, whereas the vocabulary training also included a training exercise – people made a sample plan for an example comment, and were able to see our solution for the same comment once they submitted their plan. This interactivity in vocabulary training might have caused the training task to seem less rushed, since people progressed through it with explicit feedback from us. For the planning task, there were two significant differences out of the five NASA-TLX questions: (i) people felt that they had to do significantly more hard work for making unstructured plans than vocabulary-based plans ($p < 0.05$), and (ii) people felt significantly more insecure and annoyed when making unstructured plans ($p < 0.01$). These results indicate that vocabulary-based planning requires either comparable or less effort than unstructured planning.

Plan Quality

We use three measures for plan quality: the number of steps generated per plan (granularity), the number of steps that are mechanical (atomicity), and the number of steps that can be crowd- or self-sourced (sourcing potential).

Granularity. People make unstructured plans with 2 steps and vocabulary-based plans with 2.5 steps, on average (s.d. unstructured=1, vocabulary=1). The median granularity for plans is 2 steps for both conditions. The results of a linear mixed effects model indicate that vocabulary-based plans are significantly more granular, i.e., have significantly more steps than unstructured plans ($p < 0.005$). The model uses granularity as the dependent variable, unstructured vs. vocabulary condition as the independent variable (fixed effects) and participants as the random effects.

Atomicity and Sourcing Potential. We qualitatively code

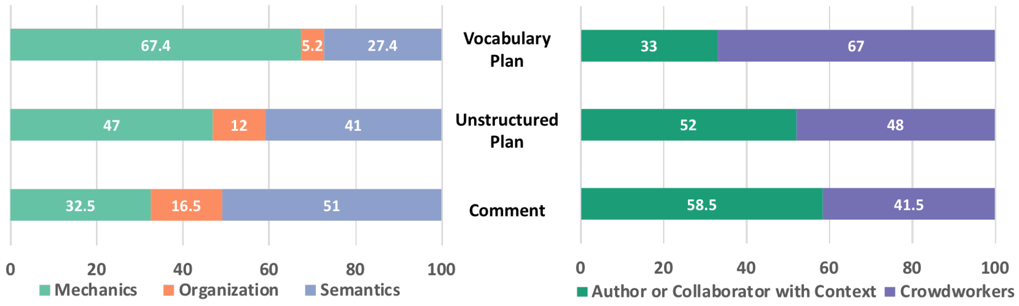


Figure 4: The percentage of comments and steps of unstructured and vocabulary-based plans that: *(left)* are mechanical, organizational or semantic in nature – higher percentage of mechanical comments implies higher atomicity; *(right)* need to be completed by author or collaborators (people with context) vs. crowdworkers (people with no context) – higher percentage of latter implies greater sourcing potential.

a random sample of 200 comments, unstructured and vocabulary-based plans for: (i) whether each comment or step is related to mechanics, organization or semantics (for atomicity), and (ii) whether each comment or step can be completed only by someone with context about the document (e.g., author or collaborator) or also by a crowdworker with minimal-to-no context. Coding is done by the first and fifth author; inter-rater reliability is calculated using additional 20 comments and plans – Cohen's Kappa is 0.77 for atomicity and 0.84 for sourcing potential (significant agreement). There is a $\sim$35% increase in mechanical steps between vocabulary-based plans and comments, and a $\sim$20% increase between vocabulary-based and unstructured plans (values depict absolute improvement). For sourcing potential, we see a $\sim$26% increase between comment and vocabulary-based planning, and a $\sim$19% increase between unstructured and vocabulary-based planning (see Figure 4). This implies that having a vocabulary when planning leads to planners using more basic steps – the author need only accomplish a small percentage of the steps in a vocabulary-based plan, the rest can be crowdsourced.

Conclusion

In this paper, we explore a mechanism for outsourcing task planning for context-embedded tasks to people who have context (e.g., collaborators), while reducing the time and effort costs and improving plan quality. We use a data-driven process to develop a vocabulary of 18 basic functions that can be used to create plans. Compared to unstructured plans that are made without any aid, we find that plans created using our vocabulary require less time and effort to be created, have more atomic steps and more steps that could be assigned to other people, thereby reducing the burden on the person who is tasked with executing the plan.

REFERENCES

1. Elena Agapie, Lucas Colusso, Sean A Munson, and Gary Hsieh. Plansourcing: Generating behavior change plans with friends and crowds. In *CSCW 2016*.
2. Peter M Gollwitzer and John A Bargh. 1996. *The psychology of action: Linking cognition and motivation to behavior*. Guilford Press.
3. Nick Greer, Jaime Teevan, and Shamsi T Iqbal. 2016. *An introduction to technological support for writing*. Technical Report. Microsoft Research.
4. Nicolas Kokkalis, Thomas Köhn, Johannes Huebner, Moontae Lee, Florian Schulze, and Scott R Klemmer. 2013. Taskgenies: Automatically providing action plans helps people complete tasks. *ACM TOCHI* (2013).
5. Jaime Teevan, Daniel J Liebling, and Walter S Lasecki. Selfsourcing personal tasks. In *CHI 2014 EA*.
6. Jeffrey M Zacks, Barbara Tversky, and Gowri Iyer. 2001. Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General* 130, 1 (2001), 29.