MS-LaTTE: A Dataset of Where and When To-do Tasks are Completed

Sujay Kumar Jauhar, Nirupama Chandrasekaran, Michael Gamon, Ryen W. White

Microsoft Research
Redmond, WA, USA
{sjauhar, niruc, mgamon, ryenw}@microsoft.com

Abstract

Tasks are a fundamental unit of work in the daily lives of people, who are increasingly using digital means to keep track of, organize, triage, and act on them. These digital tools – such as task management applications – provide a unique opportunity to study and understand tasks and their connection to the real world, and through intelligent assistance, help people be more productive. By logging signals such as text, timestamp information, and social connectivity graphs, an increasingly rich and detailed picture of how tasks are created and organized, what makes them important, and who acts on them, can be progressively developed. Yet the context around actual task completion remains fuzzy, due to the basic disconnect between actions taken in the real world and telemetry recorded in the digital world. Thus, in this paper we compile and release a novel, real-life, large-scale dataset called MS-LaTTE that captures two core aspects of the context surrounding task completion: location and time. We describe our annotation framework and conduct a number of analyses on the data that were collected, demonstrating that it captures intuitive contextual properties for common tasks. Finally, we test the dataset on the two problems of predicting spatial and temporal task co-occurrence, concluding that predictors for co-location and co-time are both learnable, with a BERT fine-tuned model outperforming several other baselines. The MS-LaTTE dataset provides an opportunity to tackle many new modeling challenges in contextual task understanding and we hope that its release will spur future research in task intelligence more broadly.

1. Introduction

Tasks are the primary unit of personal and professional productivity. People capture, organize, track and complete tasks as a way to measure and make progress towards their goals. Task management strategies range from scribbled sticky notes on refrigerators to complex group collaboration platforms such as Microsoft Planner, and everything in between. Digital tools for task management support are prevalent in a range of applications including electronic mail (Bellotti et al., 2003), to-do applications (Bellotti et al., 2004) and digital assistants (Graus et al., 2016), and for a range of user scenarios such as contextual reminders (Kamar and Horvitz, 2011), task duration estimation (White and Hassan Awadallah, 2019), and complex task decomposition (Zhang et al., 2021b). In short, digital task management provides an important opportunity to make users more productive, and help them balance their priorities in a connected world where the boundaries between life and work are often blurred.

Many applications record and use several signals such as the text of a task, the time it was created, its importance, due-date, and who it is assigned to, in order to build intelligent solutions for the user. However, they all notably assume tasks are homogeneous from the perspective of the *completion* context. This assumption is flawed because tasks are, in fact, deeply context dependent; completing or making progress on tasks depends on extraneous factors, such as time in schedule, proximity to home or businesses, and resource availability. For example, a user is likely to want a reminder to buy eggs when they are close to a grocery store, as opposed to when they are at an airport.

Two types of context that are especially salient are location and time. These signals are readily accessible to systems and have been studied in previous work on contextual understanding (Graus et al., 2016; Bellotti et al., 2004; Benetka et al., 2019), recommendation (Zhuang et al., 2011; Yao et al., 2015; Zeng et al., 2016) and reminders (Kamar and Horvitz, 2011). However, there continues to remain a disconnect between logged contextual signals and actions taken in the real world, since users often record completion of tasks at a later time and a different location (Zhang et al., 2021a). Moreover, most prior work leverages proprietary data and does not make these available to other researchers for further study. The lack of a sizeable publicly available task dataset, and especially one that is tagged with location and time meta-data, has limited research on task intelligence in general, and particularly on the important area of contextual task modeling.

Thus, in this paper we are releasing a novel resource called the *Microsoft Locations and Times of Task Execution* (MS-LaTTE) dataset. This dataset of 10,101 tasks sourced from real-world data is the largest publicly available repository of to-do tasks of any kind, and an order of magnitude larger than previously collected datasets (Landes and Eugenio, 2018). Additionally, it is the only dataset that also contains contextual location and time labels for where and when the tasks are likely to be completed. The dataset is available here under a permissive Community Data Licence Agreement that does not impose any restrictions on its use, modification or sharing.

In addition to collecting and releasing the data, we also explore MS-LaTTE in this paper, to understand its utility for contextual task modeling. Specifically, we analyze the annotations to see if they capture interesting properties or regularities that might be useful for downstream modeling. Additionally, we motivate and establish two new benchmark evaluation tasks derived from the MS-LaTTE dataset – co-location prediction and co-time prediction – and evaluate several popular language modeling approaches against these benchmarks. We find that learning is possible on both benchmarks, and a fine-tuned BERT approach significantly outperforms other baselines. However, both co-location prediction and co-time prediction are challenging problems, and allow for more complex modeling efforts in future work to outperform the systems that we evaluate. Beyond its applications to the general area of contextual task modeling, the release of this dataset has the potential to benefit modeling efforts in concrete application scenarios in task intelligence, including spatio-temporally relevant reminders and notifications, context-aware calendaring for individuals as well as groups, and focus-time scheduling for work and life. Additionally, we believe that this task-oriented dataset could be interesting to the broader NLP community more generally, as a supplemental domain centered on short texts that often lack explicit context. Other examples of this include social media posts (e.g. Twitter), utterances with a language-enabled virtual assistance, and queries issued to a search engine – all of which have been central to a broad range of research in NLP. In summary, we make the following contributions:

1. Highlight the opportunities and challenges in contextual task modeling.

2. Compile and publicly release a novel dataset with over 10k tasks (an order of magnitude larger than previously available datasets) labeled with location and time meta-data (§3).

3. Perform a detailed analysis of the location and time annotations, demonstrating that they capture intuitively reasonable task regularities (§4).

4. Conduct experiments to show the viability of the data for machine learning applications, setting up the novel benchmark tasks of co-location prediction and co-time prediction in the process (§5).

5. Present future directions, including further applications and additional contexts (§6).

2. Related Work

Previous related research includes work on task management, context-aware computing, and tasks data.

2.1. Task Management

Task management systems that assist people in successfully managing, prioritizing, structuring, organizing and completing tasks have been the subject of a large body of research. There are many applications and virtual assistants on the market today, including Amazon Alexa, Google Assistant, Trello, and Microsoft To-Do, to name only a few. Since task management does not take place in isolation, but rather in the rich context of daily life, research on the topic is situated within the broader area of context-aware computing. Location and time are two important contextual factors for both personal task assistance (Benetka et al., 2019; Wang and Pérez-Quiñones, 2014) and work-related task management (Anhalt et al., 2001; Rhodes, 1997). Even before the existence of smaller computing devices such as smartphones that allow location-aware computing, actions as simple as setting an alarm or time-based reminder demonstrate a strong need for temporal context in task management.

Bellotti et al. (2004) point out that co-location of tasks minimizes the need for multiple excursions for a user and that certain kinds of tasks show effects of periodicity, in terms of certain days, weeks and time of year. Graus et al. (2016) study reminder data from the Cortana personal assistant. They establish that different task types cluster around different user-selected notification times. Communication task reminders, for example, tend to have a notification time that falls into typical work hours (when others are likely to be available), chore notifications are more commonly set to early morning hours and tasks that involve moving to a specific location tend to have reminders set for lunch time at work. Interestingly, in their data set, knowing the time at which a reminder was created is a stronger predictor of notification time than the task title itself. Ludford Finnerty et al. (2006) show that both location and movement patterns are important in reminder applications. By combining location and time information, context-aware task management systems can better assist users by recommending the right tasks at the right times (Kessell and Chan, 2006; Rhodes, 1997).

2.2. Context-Aware Computing

Beyond task management, in context-aware computing and recommendation systems, time and location signals are also fundamental. Context-aware recommendation of points of interest (POI) is typically studied based on check-in data from social network applications such as Foursquare and uses location as well as time (Yao et al., 2015; Zhao et al., 2019). POI recommendation also has a strong social component, making it amenable to the use of collaborative filtering signals in addition to spatio-temporal context (Yuan et al., 2013). Other applications include generating online recommendations for advertising and news, based on temporal patterns (Zeng et al., 2016), and recommending entities based on historic user behavior and spatio-temporal sensor context (Zhuang et al., 2011).

Other broad research areas where time and location (as well as social context and travel trajectory) provide important information are mobile search (Teevan et al., 2011; Amini et al., 2012), opportunistic routing (efficient diversions to vehicular driving paths) (Horvitz

and Krumm, 2012), and in information retrieval in general (Bennett et al., 2011; Radinsky et al., 2013).

2.3. Tasks Data

Despite the extensive research in task management and assistance, there is very little publicly available data. Typically, the data sources used in the literature are proprietary and subject to strict privacy requirements. To the best of our knowledge, Landes and Eugenio (2018) are the only researchers to have publicly released a dataset of to-do tasks. It consists of approximately 600 tasks provided by users of Trello who volunteered their data, as well as from public Trello boards. The tasks are annotated for (a) an intelligent agent that would be appropriate to assist with the task (the set of 15 agents serve as a task taxonomy), and (b) the argument(s) of the action expressed in the task.

In comparison, MS-LaTTE is the first public large-scale dataset of real-world to-do tasks, consisting of more than 10k instances. It is also the only dataset to contain annotated meta-data about the locations and times at which tasks are likely to be completed.

3. Dataset Collection

In this section, we describe the annotation setup for collecting the MS-LaTTE dataset. Our goal is to source real-world tasks and capture the locations and times at which they are usually completed.

3.1. Sourcing Real-world To-do Data

We source tasks for our dataset from a sample of the logs of the now-defunct Wunderlist application. These logs were obtained after a thorough legal- and trust-approved enterprise-grade pipeline has processed them to anonymize and scrub all personally identifiable information. In addition, the pipeline performed k-anonymization so that tasks that were created by fewer than five users or fewer than 100 times in total are automatically discarded. Finally, named entities recognized in text are deterministically replaced with a different name, in order to maintain privacy and anonymity. The result is an *aggregate* view of the logs from English, French, Italian, German and Spanish locales devoid of any identifiers, private information, or infrequent tasks that can be correlated back to an individual user. What remains is a collection of task titles (such as "buy milk", "mow the lawn", etc.) along with list titles to which they are commonly assigned by users (such as "groceries", "home chores", etc.).

It is important to note that, due to our rigorous privacy-preserving anonymization process, list titles are only nominally lists, since it is impossible to recover an actual user list consisting of multiple tasks. Hence the task title is the central unit of the aggregate logs, and the list title can be considered to be just an associated piece of meta-data for the task.

Some basic text processing is then applied to the aggregate logs. Titles that have non-ISO-8859-1 characters are discarded; those that remain are lowercased, their punctuation is removed, and their accented characters are replaced with non-accent equivalents.

This aggregate, processed view provides a rich collection of real-world to-dos from which we sample items for annotation. Unfortunately, grocery – and to a lesser extent packing – related tasks are over-represented in the aggregate data, accounting for over 70% of distinct items, by our estimates. Therefore,we use heuristics to under-sample grocery and packing tasks. This is to avoid an annotated dataset where a large part is trivially assigned a single location (or time) label. Our heuristic uses a manually curated set of the most popular list titles from the data related to groceries and packing (e.g. "grocery", "safeway", "packing list"), and samples task titles from these lists at a lower rate than from other lists. Specifically we use an 10%, 10%, 80% sampling probability for grocery, packing, and other lists respectively. Additionally, we use a filter to only sample tasks in the aggregate data that are sourced from an English locale.

In this manner, we sample a total of 12,000 distinct task-list pairs that are subsequently annotated with location and time information.

3.2. Annotation Setup

We used an internal crowd-sourcing platform to delegate work to non-expert workers in India contracted to perform annotations. These workers were paid a fixed hourly rate of $11.98 USD rather than an amount per HIT completed, which in our experience leads to better annotation quality. We conducted the annotation over the sampled collection of task-list pairs in two distinct stages: first for location, then for time. This ordering was motivated by a couple of initial pilot annotation rounds where we observed that annotators agreed on location to a much higher degree than on time; this makes sense, since tasks are much more likely to be completed at the same location by different users than at the same time (due to individual schedules and preferences).

Annotations were collected between May and August 2020, during the COVID-19 pandemic, when boundaries between life and work were especially blurred, and activities as well as schedules were atypical due to lockdowns, social distancing and other precautionary behavior. To account for this, annotators were instructed to use pre-COVID times as a contextual frame of reference. Annotators were presented with an interface like the one in Figure 1, where the unit for a HIT was a single task-list pair. After providing their consent to have their annotations collected for research and machine learning, their judgments were collected across the two different stages.

In the first stage, annotators were asked **WHERE** they would normally complete a task and were provided with four broad location categories from which to select one or more: (a) Home (b) Work (c) A public location (d) Somewhere else (along with a free-form text

Figure 1: The interface used for collecting location and time for task completion. Question 2 differs between stages of annotation; in this example an annotator is asked to provide time labels

box). Note that our guidelines broadly specify that "Work" can cover a number of locations, depending on a task creator's likely primary vocation, including "school" or "college" for students. Further, if (c) is selected, annotators were asked to specify at least one of several public location labels (e.g., "grocery store," "dentist," "library," etc.). Initially, a list of 36 such public locations were manually curated from a taxonomy of map location categories. Then, over a few initial pilot rounds of annotation, inputs from option (d) were used to refine this list into a final set of 69 public location labels. For annotator convenience, public locations were manually organized into seven broad categories such as "retail", "recreation", "finance", etc. – these were only shown in the annotation interface, and do not form part of the final dataset. Notably, annotators were instructed to respond by providing labels for *physical* locations of task completion. That is, even if a task may be completed online, the labels reflect the physical locations at which it is normally completed. This was done to reduce confusion and ambiguity, as well as to make the data more directly grounded in the real world and usable by future applications that can leverage geo-location information. Tasks that are completed online are nevertheless an interesting area of research for future work.

The full set of 12,000 HITs were labeled by three annotators in the first round. Those HITs that were marked as unfamiliar or unknown to two or more annotators were discarded from the dataset. Furthermore, in instances where all three annotators did not have a single location label in common, the first author of this paper acted as a fourth annotator; 365 HITs were thus supplementally annotated. The few remaining instances where there was four-way disagreement on location labels were also subsequently discarded. In the end, a total of 1,899 task-list pairs from the original set of

12,000 were removed, leaving 10,101 tasks.

In the second round, only the remaining tasks were annotated by asking judges **WHEN** they would normally complete these tasks. They were asked to select one or more time labels from a set of 10 time buckets, (as shown in Figure 1) each specified by two distinct dimensions: (1) the time of the day, and (2) the day of the week. Times of the day included: (a) Morning (b) Afternoon (c) Evening (d) Night, and (e) Anytime; while days of the week could be either: (i) Weekday, or (ii) Weekend. Annotators were instructed to interpret the different times of day according to their own frame of reference (for example *Morning* might mean 5am-8am to an early riser, but 8am-11am for someone else), thereby allowing for label alignment across annotators and tasks at a conceptual level rather than according to strictly defined time buckets. Additionally, they were told to use *Anytime* when they were no more likely to complete a task at any one of the other times of day. Annotators were not instructed to account for expected task duration, and for the purposes of this study we did not distinguish between shorter, simpler tasks or longer, more complex ones.

Since when someone completes a task is altogether more subjective than where, it is expected that there will be a great deal more variation in labels for this second stage of annotation. To account for and capture some of this variability we asked five annotators to label each of the 10,101 task-list pairs in the dataset, and none of the responses were discarded. Of course, even five annotations is insufficient to fully capture the breadth of preferences for when tasks are completed; however, the dataset we have collected allows for easy extension of time information with more responses in future work.

In summary, the final MS-LaTTE dataset – collected over two stages of annotation – consists of 10,101 real-

world task-list pairs, each of which includes a set of labels for location from three annotators[1], and for time from five annotators. To the best of our knowledge, it is the largest (by an order of magnitude) dataset of to-do like tasks sourced from real users, and the only one to contain explicit contextual information in the form of locations and times at which tasks are completed.

4. Analysis of the Dataset

In this section, we describe an analysis of MS-LaTTE. We begin by measuring agreement between annotators. We then investigate the distributions of labels across location and time independently, and what annotator consensus looks like for each. Finally, we examine the relationship between location and time labels by performing a cross-correlation exploration of the data.

4.1. Annotator Agreement

We use Krippendorff's Alpha (Krippendorff, 2011) to measure the degree of annotator agreement. Since annotators can provide multiple labels to instances in the dataset, we apply the MASI (Passonneau, 2006) distance metric, a measure of agreement over set-valued objects. This setup yields agreement values of 0.50 and 0.09 for Location and Time respectively, which in turn are considered moderate and poor degrees of inter-rater agreement (McHugh, 2012).

As noted in §3.2, we hypothesized that labels on Time are subjective and therefore the low annotator agreement is expected. The ability for annotators to provide multiple labels for each instance also negatively impacts the inter-rater reliability metric. If singleton labels – that is, those that were provided only by one annotator for a given instance – are removed, Krippendorff's Alpha values become 0.87 and 0.26, which are excellent and fair for location and time respectively. In other words, while human judges may have individual preferences for where or when they expect to complete tasks, they are more likely to agree on a core set of locations or times at which these tasks should be completed. This is important, not only to contextualize the agreement numbers, but because it means that we may be able to leverage the existence of this core agreement set for predictive modeling.

4.2. Label Distribution

We now turn to a snapshot view of the label distributions over location and time portions of the data, in the form of histograms. They are shown in Figures 2a and 2b respectively. Note that log counts (base 2) are used instead of raw counts to show differences between labels more clearly. We plot the histogram over labels on which there was majority agreement between annotators; this is to mitigate any impact that singleton labels (as noted in §4.1) may have on observable trends.

Figure 2a demonstrates that *home* and *work* are by far the most popular locations for tasks in the dataset, and are an order of magnitude more frequent than any of the other labels. *Grocery* stores are the third most frequent label, despite the heuristics that we used to undersample grocery related tasks (§3.1). The rest of the labels, corresponding to purchase or errand related locations (and including the remaining 60 not shown on the figure), follow a long-tailed distribution. These general trends align with our observations about the larger aggregate Wunderlist logs, which forms the basis for the MS-LaTTE dataset; namely that grocery, home, and work-related tasks formed the overwhelming majority of tasks that users tracked in Wunderlist (and perhaps also in task management applications in general).
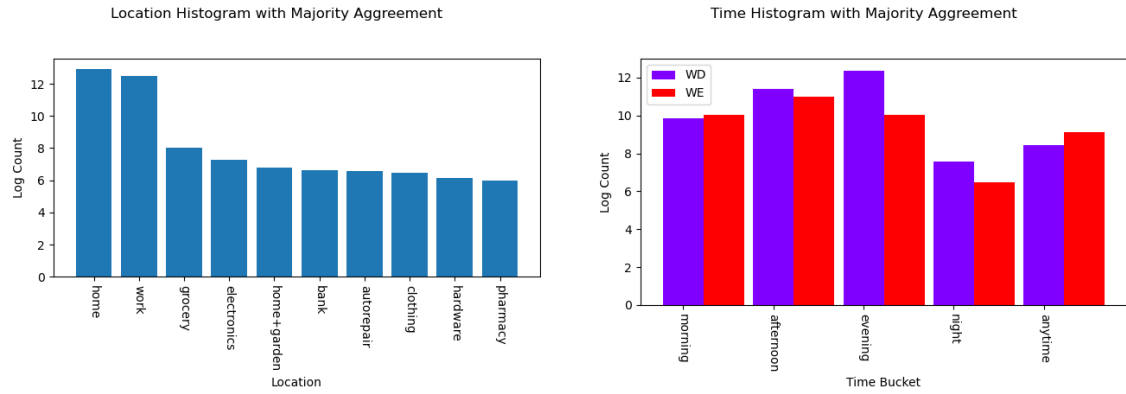
Meanwhile, Figure 2b also has some interesting trends over time label distribution. Weekday evenings appear to be the most active time for completing tasks, perhaps due to the likelihood of users possibly being at *home*, *work*, or running errands (like at a *grocery* store), where many tasks are frequently completed. The popularity of *home* or *work* related tasks is also potentially a contributing factor to why mornings and afternoons are additional common times for completing tasks. A more in-depth analysis on cross-correlations between location and time labels is presented in §4.4.

The difference between weekdays (WD) and weekends (WE) shows some interesting properties too. While people are generally less active completing tasks on weekends (sometimes starkly, such as in the evenings), they are actually slightly more active on weekend mornings – perhaps due to home chores or other errands that are set aside for non-working days. Additionally, tasks that are categorized as being done *anytime* (often very short, simple tasks that require little planning, as we show in §4.3) are more likely to be completed on weekends than on weekdays. This is perhaps due to weekends providing more leisure time for unplanned tasks.

4.3. Label Consensus

Our computation of annotator agreement by discarding singleton labels (§4.1), seemed to indicate that annotators tend to agree on core sets of labels for specific tasks, especially in the case of location. Thus, we now provide examples of high-agreement labels for both location and time, to show that the information provided by annotators captures reasonable location and temporal expectations for task completion. Tables 1 and 2 provide some examples of tasks that were assigned the same label by a majority of annotators; not all labels are included in either table due to space constraints.

Table 1 demonstrates that location labels are often incontrovertible and that annotators are able to correctly agree on the most likely location for tasks, despite the more than 70 possible labels to choose from. Meanwhile, Table 2 also shows that when the majority of annotators agree on a time bucket for a task, they select

---

[1]222 instances in the dataset contain judgments from four rather than three annotators, for the reasons noted above.

(a) A histogram (counts in log base 2) of the 10 most popular location labels, where the majority of annotators agreed on the location label.

(b) A histogram (counts in log base 2) of time bucket labels, where the majority of annotators agreed on the time bucket label. WD=weekday and WE=weekend.

Figure 2: Histograms of majority agreement labels over the Location and Time annotations in the dataset. The y-axes are presented in log (base 2) of the raw counts to accentuate differences between labels.

home	work	office supply	pharmacy	electronics store	clothing store	hardware store
rearrange closet	meeting tasks	buy envelopes	dr refill	bring in headphone	office attire	get a tape measure
fix tv remote	sociology paper	buy sharpies	zzzquil liquid	dad speakers	astronaut costume	pressure washer part
put on license sticker	finish udemy course	get packing tape	pick up relpax	more usb cables	scouts uniform	make house key copy

Table 1: Examples of tasks that annotators commonly agreed should be assigned to specific location categories.

WD morning	WD afternoon	WD evening	WD anytime	WE morning	WE afternoon	WE evening
antibiotic morning	meeting tasks	bring book in	call sumo	antibiotic morning	platinum caulk	return items
live class	sort out direct debits	pickup drycleaner	phone to mummy	finish gym	paint toilet	target vitamins for week
get breakfast and coffee	work on newsletter	upload instagram	email to pronob	water poinsettias	lunch with parents	change tank filter

Table 2: Examples of tasks that annotators commonly agreed should be completed during a specific time bucket.

labels that are very reasonable. In other words, while individuals may choose to complete some of these example tasks at different times, due to personal preferences or schedules, when several judges agree on a time bucket for a task, it appears to be a label that is easily interpretable – e.g., work tasks on weekday afternoons, or errands and hobbies in the evenings. Additionally, as previously noted, the *anytime* label is often associated with very short tasks that do not require prior planning. While Tables 1 and 2 provide agreement on single labels, a reasonable follow-up question is to ask whether multi-label annotations also capture useful signal. We attempt to answer this question by calculating the point-wise mutual information (PMI) between location and time labels independently, when these labels are applied to the same task by annotators. Intuitively, the PMI captures the degree to which labels are co-assigned, thereby providing a way to assess the interpretability of multi-label annotations.

Table 3 provides examples of such co-assignments. Each column gives the top three labels (by PMI) that co-occur with the label in the column header. Note that items within a column are unrelated, and their only shared connection is co-occurrence with the column header. Only a subset of location and time labels are provided, for brevity. As can be seen from the table, co-assignment by annotators often makes intuitive sense. For example, laundry can be done both at home or at

a laundromat, home & garden stores sell similar goods to hardware stores, grocery stores and pharmacies are co-located in the same building, and short unplanned tasks can be done anytime on weekdays or weekends.

4.4. Location and Time Cross-Correlation

Up to this point, our analyses have considered the location and time parts of the dataset separately, but since each of the 10,101 tasks in the dataset are annotated with both types of labels, we can also conduct an analysis that looks at them jointly. Specifically, we can investigate whether location labels assigned to tasks make intuitive sense with respect to time, and vice-versa. Stack plots attempting to answer the question of cross-correlation between label sets are presented in Figures 3a and 3b.

For each pair of location category and time bucket, we count the number of tasks that were assigned both labels (each by majority agreement). Then we marginalize over time and location to get the stacked bars in Figures 3a and 3b, respectively. Note that we only use the 10 most popular location labels in this analysis.

These plots reveal some interesting and intuitively sensible findings. For example, in Figure 3a, *home* and *work* are the only two locations where tasks are completed in every time bucket[2]; this seems reasonable

---

[2]It may be surprising that a number of *work* tasks are com-

		Location		Time		
home	home+garden store	grocery store	doctor's office	WD morning	WE afternoon	WD anytime
laundromat	hardware store	pharmacy	hospital	WD afternoon	WE evening	WE anytime
beauty salon	sporting goods store	home+garden store	pharmacy	WD evening	WD evening	WD afternoon
work	grocery store	home	home	WE morning	WE morning	WD morning

Table 3: Examples of location and time labels that were commonly co-assigned by annotators, as measured by point-wise mutual information.
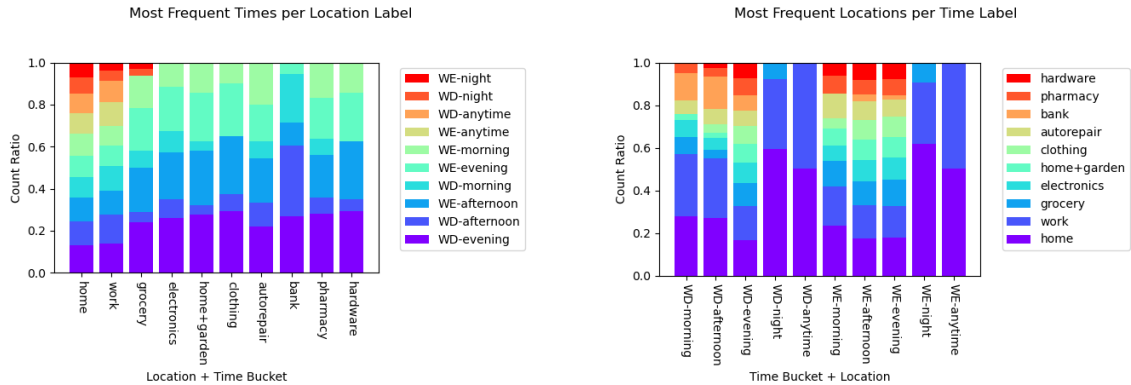
(a) The distribution of time bucket labels over the 10 most popular location categories.

(b) The distribution of the 10 most popular location labels over time buckets.

Figure 3: Stack plots showing the distribution over Location and Time label pairs.

since people spend the most time at these two locations. Another interesting observation is the relative proportion of tasks done on weekends at different public locations. For example, far more *grocery*, *home & garden* or *hardware* tasks are completed on weekends than *bank* tasks, since it is the expectation that the former are open for business while the latter are not (or may only be open with limited hours).

Figure 3b also contains some noteworthy details. One example is that tasks assigned *anytime* labels are only done at *home* or at *work*; this makes a lot of sense, considering that they are typically short, unplanned ones (§4.3), and that tasks completed at any other public location typically requires some form of forethought or planning. Another example, is the relative proportion of *work* related tasks completed on different days of the week (especially in the mornings and afternoons), which conforms to the expectation that most people work on weekdays, rather than on weekends.

Figures 3a and 3b present a broad picture of cross-correlation between location and time labels, but it may be useful to also explore a more focused view of cross-correlation. Table 4 presents some examples of the most highly correlated pairs of labels, as measured by PMI. This value is computed from probabilities of independent and joint label occurrences, which are obtained from counts over the location and time labels that received majority agreement. Because PMI is sensitive to very infrequent events, we discarded location labels that have fewer than five tasks associated with

Location	Time
restaurant	WD+WE night
dmv	WD afternoon
movie theater	WE evening
library	WD+WE night
car wash	WE morning
gym	WE morning

Table 4: Examples of some location and time labels that were highly correlated, as measured by point-wise mutual information.

them. As can be seen from the table, these highly correlated label pairs make intuitive sense: for example, the fact that *restaurant* and *library* are associated with the *night* time bucket on both weekdays and weekends (i.e., date nights or study sessions), that *gym* is associated with *WE morning* (i.e., early workout on weekends), or that *dmv* is associated with *WD afternoon* (i.e., less busy times when people are often at work[3])

In summary, our main takeaway from the dataset analysis is the fact that while the data does contain some expected variance – in the form of individual latitude for times at which tasks are completed – interesting and often intuitively reasonable properties of task completion are captured by annotations with majority agreement. This motivates the use of MS-LaTTE for the learning effort we tackle in §5, as well as for future work that we hope will leverage the dataset for modeling task intelligence.

pleted on weekends, but recall this label is vocation dependent, and includes locations such as college (§3.2).

[3]See: helpful hints at www.ncdot.gov

5. Modeling Co-Location and Co-Time

Given the annotations in the dataset, there are many interesting predictive problems that can be tackled such as predicting the location or time bucket most likely to support some task activity, or predicting when (resp. where) a task should happen given a user's set location (resp. time bucket) and description, or even scheduling a user's day using their time commitments and likely locations. However, in this paper we seek only to describe and validate the MS-LaTTE dataset and therefore tackle the two foundational modeling problems of predicting whether two tasks are likely to be completed together (by location or time); other modeling efforts are left to the community and to future work.

Note that while these tasks of co-location and co-time prediction may be simpler than some of the more ambitious modeling challenges described above, a model that successfully tackles the former may indicate approaches that might succeed on the latter. Moreover, co-location prediction and co-time prediction are meaningful modeling efforts by themselves, leading to potential user-facing scenarios such as alerting users working on a given task, to other tasks they can complete, based on their current location or time (or both). Formally given two tasks $\mathcal{T}_1 = (t_1, l_1)$ and $\mathcal{T}_2 = (t_2, l_2)$, where t and l are task and list descriptions respectively, the problems of co-location and co-time are to find binary predictive functions $f(\mathcal{T}_1, \mathcal{T}_2) \rightarrow 0, 1$. We generate benchmark datasets for evaluating these two predictive tasks from the annotated MS-LaTTE dataset. We sample 25,000 task pairs from the cross product of the 10,101 unique tasks in MS-LaTTE, using 20,000, 1,000, and 4,000 respectively for training, validation, and test splits. Pairs are assigned a positive label if they contain at least one common label that was assigned by a majority of annotators; otherwise they are assigned a negative label.

While pairs are unique, tasks themselves may repeat. To ensure fair evaluation we stratify the dataset so that tasks that appear in any split do not appear in any other split. It may be noted that the resulting benchmark datasets are imbalanced containing roughly 71/29 and 38/62 positive/negative splits for location and time respectively. The train, validation, and test splits for co-location and co-time benchmarks are released with the dataset. In this paper, we report accuracy and Macro F1 (due to the imbalance in the datasets) as evaluation metrics for the two evaluation tasks.

We evaluate several popular language modeling approaches in this paper. They include: (a) **Random** – a baseline that randomly assigns a positive or negative label to an instance, using the ratios in the training data as bases for sampling a label. (b) **Lexical** – a model that featurizes both sets of tasks and list strings and uses uni-, bi- and tri-gram features in a logistic regression classifier trained and tuned on the train and validation splits respectively. (c) **GloVe** – a model that uses the popular GloVe vectors (Pennington et al., 2014) as features for lexical items in the task and list strings. An average of the GloVe vectors is passed through a logistic classifier that is trained and tuned on the train and validation splits respectively. (d) **BERT** – a model that is similar to the GloVe model above, but uses pre-trained BERT embeddings (Vaswani et al., 2017) instead. Because these representations capture full strings as opposed to individual tokens, the embeddings for $\mathcal{T}_1$ and $\mathcal{T}_2$ are concatenated rather than averaged.

Finally, we compare these models, against a more sophisticated fine-tuned BERT model. Like the pre-trained BERT model, this one also featurizes both task inputs $\mathcal{T}_1$ and $\mathcal{T}_2$. But it uses a composition variant that has been successfully applied in past work to problems that involve dual string comparison (Mou et al., 2015), such as textual entailment:

$$c = [\mathcal{T}_1; \mathcal{T}_2; \mathcal{T}_1 - \mathcal{T}_2; \mathcal{T}_1 \circ \mathcal{T}_2] \qquad (1)$$

where the semi-colon represents concatenation, and the $\circ$ signifies the element-wise dot product. In our model (**BERT TE-FT**) the representation c is then passed through a non-linear layer, before a final linear layer with a sigmoidal output produces a binary prediction value. The model uses an intermediate dimension of 256 as an output to a ReLU non-linearity; it also applies a dropout factor of 0.5 to the model during training to avoid overfitting. We use binary cross-entropy as the loss function, and we train the model using an Adam optimizer with a fixed weight decay (Loshchilov and Hutter, 2017), and initial hyperparameters of $lr = 1e^{-5}, eps = 1e^{-8}$.

We note that none of the models above are especially sophisticated or complex; this is by design. Our evaluation (and the models' purpose) is designed to establish a benchmark for the two novel tasks, by showcasing their tractability, while leaving room for future work to improve upon our results.

	Co-location		Co-time	
	Accuracy (%)	Macro F1	Accuracy (%)	Macro F1
Random	59.00	0.501	51.85	0.492
Lexical	77.38	0.696	57.47	0.518
GloVe	74.48	0.613	**60.40**	0.498
BERT	75.40	0.625	59.28	0.511
BERT TE-FT	**81.98**	**0.773**	59.10	**0.569**

Table 5: Results of several models on the binary prediction tasks of co-location and co-time detection.

The results of our evaluation are given in Table 5. They show that while all models outperform the random baseline, they do so with varying degrees of success. The best model is the BERT TE-FT model, which outperforms the other approaches by a significant margin[4] on both benchmark tasks.[5] The lexical model proves surprisingly capable, outperforming both GloVe and

---

[4]Statistically significant at a p-value=0.01 based on a paired student's t-test.

[5]It does not achieve the highest accuracy on co-time prediction due to the imbalanced nature of the dataset.

BERT models; this is possibly because the training data size is fairly large and contains a decent coverage of lexical terms. Notably, both tasks are challenging (co-time prediction being significantly more so), and there is room for improvement over the simple approaches described in this paper.

	Co-location		Co-time	
	Accuracy (%)	Macro F1	Accuracy (%)	Macro F1
BERT TE-FT	**81.98**	**0.773**	59.10	**0.569**
(-) Fine-tuning	71.58	0.426	**61.45**	0.381
(-) TE Composition	80.90	0.757	56.70	0.541

Table 6: An ablation study demonstrating the effects of removing components of the full BERT TE-FT model.

While BERT TE-FT is clearly the best model on both benchmark tasks, we can also evaluate how much contribution its features add by conducting an ablation study. Specifically, we ablate the model by: (a) freezing the parameters of the BERT model (effectively negating the effects of fine-tuning); or (b) using a simple concatenation of $\mathcal{T}_1$ and $\mathcal{T}_2$ instead of the more complex vector composition variant in Equation 1. The results of our ablation study are given in Table 6. They show that while both components influence the full model positively, fine-tuning is clearly a far more important positive factor. This seems to indicate that the language used in to-do tasks is different from general purpose text on which BERT is trained, and thus benefits from fine-tuning to the domain.

In conclusion, we presented the two new benchmark tasks of co-location and co-time prediction, derived from the MS-LaTTE dataset. We compared several popular language modeling approaches on these benchmarks and showed that they can indeed be modeled successfully. Notably, a model containing fine-tuned BERT seemed to perform best. However, both benchmarks are challenging and present future work with interesting possibilities to outperform the simple models in this paper with more sophisticated approaches.

6. Conclusions and Future Work

We have publicly released a new dataset of to-do tasks called MS-LaTTE. This dataset contains location and time labels from multiple annotators for every one of its 10,101 tasks, and is the first to contain such contextual information surrounding task completion. It is also the largest publicly available dataset of real-world to-do tasks of any kind, by an order of magnitude. In this paper, we have described the setup used to collect and annotate the dataset, conducted a detailed analysis of its labels and properties, and performed experimental evaluations on two novel benchmark tasks – co-location prediction and co-time prediction – derived from the dataset. We found that the data captures several intuitive regularities, and that these regularities can be modeled by popular language modeling techniques – including, most successfully, by a BERT fine-tuned approach. We anticipate that the release of MS-LaTTE

will spur the research community to work on contextual task modeling, and more generally on task intelligence. Despite the utility we hope the community will derive from this dataset, it does have some limitations. We rely on third-party judges' interpretations of tasks that they did not create themselves, and for which they have very little information besides the raw textual representation. Moreover, annotators are all from a single locale (India) and thus may miss broader cultural or country-specific subtleties that are outside their field of experience. These issues may have potentially led to labels that are erroneous or noisy due to misconstrued intent. To resolve these and other issues, there are several interesting and challenging future research directions. These include: (a) alternative mechanisms to gather context information for tasks directly from individuals, such as in-situ data collection and experience sampling methods; (b) leveraging contexts beyond time and place, such as people, activity, busyness, and resources required (digital as well as physical); (c) using MS-LaTTE to model novel contextual prediction and recommendation scenarios, such as predicting when or where a task is likely to be completed, ranking tasks according to their likelihood of being completed given location, time or both, or recommending other tasks based on what a user is currently doing, where and/or when; (d) manually auditing model predictions to better understand error patterns, and devise better strategies to address them; (e) personalization of models to individuals or cohorts; (f) and integration of models in real-world task management systems and digital assistants, with user studies and online experimentation to determine the impact on users' productivity.

7. Bibliographical References

Amini, S., Brush, A., Krumm, J., Teevan, J., and Karlson, A. (2012). Trajectory-aware mobile search. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 2561–2564.

Anhalt, J., Smailagic, A., Siewiorek, D., Gemperle, F., Salber, D., Weber, S., Beck, J., and Jennings, J. (2001). Toward context-aware computing: experiences and lessons. *IEEE Intelligent Systems*, 16(3):38–46.

Bellotti, V., Ducheneaut, N., Howard, M., and Smith, I. (2003). Taking email to task: the design and evaluation of a task management centered email tool. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 345–352.

Bellotti, V., Dalal, B., Good, N., Flynn, P., Bobrow, D. G., and Ducheneaut, N. (2004). What a todo: studies of task management towards the design of a personal task list manager. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 735–742.

Benetka, J. R., Krumm, J., and Bennett, P. N. (2019). Understanding context for tasks and activities. In

Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, pages 133–142.

Bennett, P. N., Radlinski, F., White, R. W., and Yilmaz, E. (2011). Inferring and using location metadata to personalize web search. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 135–144.

Graus, D., Bennett, P. N., White, R. W., and Horvitz, E. (2016). Analyzing and predicting task reminders. In *Proceedings of the ACM UMAP Conference on User Modeling Adaptation and Personalization*, pages 7–15.

Horvitz, E. and Krumm, J. (2012). Some help on the way: opportunistic routing under uncertainty. In *Proceedings of the ACM UbiComp Conference on Ubiquitous Computing*, pages 371–380.

Kamar, E. and Horvitz, E. (2011). Jogger: models for context-sensitive reminding. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pages 1089–1090.

Kessell, A. and Chan, C. (2006). Castaway: a context-aware task management system. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 941–946.

Krippendorff, K. (2011). Computing krippendorff's alpha-reliability. Article downloaded from https://repository.upenn.edu/asc_papers/43. Last accessed 10-17-2021.

Landes, P. and Eugenio, B. M. D. (2018). A supervised approach to the interpretation of imperative to-do lists. *arXiv preprint arXiv:1806.07999*.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ludford Finnerty, P., Frankowski, D., Reily, K., Wilms, K., and Terveen, L. (2006). Because i carry my cell phone anyway: functional location-based reminder applications. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 889–898.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

Mou, L., Men, R., Li, G., Xu, Y., Zhang, L., Yan, R., and Jin, Z. (2015). Natural language inference by tree-based convolution and heuristic matching. *arXiv preprint arXiv:1512.08422*.

Passonneau, R. J. (2006). Measuring agreement on set-valued items (masi) for semantic and pragmatic annotation. In *Proceedings of the Conference on Language Resources and Evaluation*, pages 831–836.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: global vectors for word representation. In *Proceedings of the EMNLP Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Radinsky, K., Svore, K. M., Dumais, S. T., Shokouhi, M., Teevan, J., Bocharov, A., and Horvitz, E.

(2013). Behavioral dynamics on the web: learning, modeling, and prediction. *ACM Transactions on Information Systems*, 31(3):1–37.

Rhodes, B. J. (1997). The wearable remembrance agent: a system for augmented memory. *Personal Technologies*, 1(4):218–224.

Teevan, J., Karlson, A., Amini, S., Brush, A., and Krumm, J. (2011). Understanding the importance of location, time, and people in mobile local search behavior. In *Proceedings of the Conference on Human Computer Interaction with Mobile Devices and Services*, pages 77–80.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the NeurIPS Conference on Advances in Neural Information Processing Systems*, pages 5998–6008.

Wang, Y. and Pérez-Quiñones, M. A. (2014). Exploring the role of prospective memory in location-based reminders. In *Proceedings of the ACM Joint Conference on Pervasive and Ubiquitous Computing*, pages 1373–1380.

White, R. W. and Hassan Awadallah, A. (2019). Task duration estimation. In *Proceedings of the ACM WSDM Conference on Web Search and Data Mining*, pages 636–644.

Yao, L., Sheng, Q., Qin, Y., Wang, X., Shemshadi, A., and He, Q. (2015). Context-aware point-of-interest recommendation using tensor factorization with social regularization. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1007–1010.

Yuan, Q., Cong, G., Ma, Z., Sun, A., and Thalmann, N. (2013). Time-aware point-of-interest recommendation. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 363–372.

Zeng, C., Wang, Q., Mokhtari, S., and Li, T. (2016). Online context-aware recommendation with time varying multi-armed bandit. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2025–2034.

Zhang, C., Kiseleva, J., Jauhar, S. K., and White, R. W. (2021a). Grounded task prioritization with context-aware sequential ranking. *ACM Transactions on Information Systems*, 40(4):1–28.

Zhang, Y., Jauhar, S. K., Kiseleva, J., White, R. W., and Roth, D. (2021b). Learning to decompose and organize complex tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2726–2735.

Zhao, P., Zhu, H., Liu, Y., Xu, J., Li, Z., Zhuang, F., Sheng, V. S., and Zhou, X. (2019). Where to go next: a spatio-temporal gated network for next poi recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5877–5884.

Zhuang, J., Mei, T., Hoi, S. C., Xu, Y.-Q., and Li, S.

(2011). When recommendation meets mobile: contextual and personalized recommendation on the go. In *Proceedings of the ACM UbiComp Conference on Ubiquitous Computing*, pages 153–162.