Personalized Memory Testing for Names Using Email Archives

Sudheendra Hangal
Computer Science Department
Ashoka University
hangal@ashoka.edu.in

Allyson Rosen
Department of Psychiatry and Behaviour Sciences
Stanford University
rosena@stanford.edu

Ankit Mathur
University of California, Berkeley
ankitmathur@berkeley.edu

Monica S. Lam
Computer Science Department
Stanford University
lam@cs.stanford.edu

ABSTRACT

We describe a personalized and scalable system for testing of autobiographical memory (specifically, the recall of names) using email archives. We have developed a novel system that creates a personalized test by analyzing the subject's sent email messages over the past year. The system generates fill-in-the-blank questions from the user's messages, with the answers being names determined to be significant in the email corpus using text analysis. Such testing can be ecologically valid and deeply personalized compared to existing techniques. We describe technical aspects of question and answer generation and report on a study where 35 participants answered an aggregate of 1,400 questions.

We obtained a dataset of about 80 features per question and user response. Our preliminary analysis of this dataset supports several expected memory characteristics, such as that recall dips over time, and answer terms that span many days are better remembered, as are terms that are mentioned often. It also hints at some differences between memories associated with different sentiments.

Our technique can be regularly employed by individuals without incurring major expense, in the privacy of their homes. Moreover, one of our findings is that users enjoyed taking such a test (31 of 35 users signed up to take more tests in the future), implying that such tests may be a practical and effective way of testing and studying autobiographical memory on a large scale and would lead to better screening tests for memory disorders.

Keywords

Autobiographical memory, memory testing, life-logging, email, personal archives, dementia.

Categories and Subject Descriptors

H.5.m. [**Information Interfaces and Presentation (e.g. HCI)**]: Miscellaneous

1. INTRODUCTION

In this paper, we describe a system that automatically creates deeply personalized tests for an individual's autobiographical memory. Our technique utilizes the phenomenon that millions of individuals are accumulating personal digital archives which implicitly record or reflect their daily lives. Our research investigates the question of whether and how these digital footprints can be used for effective memory testing.

Current ways of studying autobiographical memory tend to be limited in timescale, and expensive to administer widely. While early research searched painstakingly for subject's personal histories that could be verified, modern approaches often they ask subjects to collect data during their everyday life. However, this process is cumbersome and must occur over limited spans of time. In contrast, we propose a technique that starts from "life-log" data that is accumulated passively and on an ongoing basis for a large section of the population. Specifically, email is an ideal data source that captures links to many autobiographical events. Long-term email archives are often available with many users, and they carry meaningful and private communication between the user and a variety of individuals. Email is used consistently by many people, on virtually a daily basis, and it is routinely archived because it is used as a tool of record. Further, every message comes with precise timestamps, which allows high resolution definition of delay, a critical variable in the fidelity of memory.

An important reason for studying autobiographical memory is that it is known to robustly activate the hippocampus, a structure of critical importance that is vulnerable to various types of brain damage such as epilepsy, oxygen deprivation, Alzheimer's disease, and aging. It is also important to characterize the differences between recent and more distant memories, as they can be differentially disrupted in some clinical conditions and therapies[15].

In clinical settings, standard cognitive assessments involve an estimate of intellectual ability and then memory, attention and executive control (mental flexibility), visuospatial and language abilities. The measures that come close to real-world activity ask patients about typical behaviors in everyday life; however, much of this information has been rehearsed over many years and is thus resilient to impairment. In contrast the ability to remember episodes is one of the more fragile types of memory. It is this type of measure

that is one of the most sensitive early warning signs of Alzheimer's Disease.

1.1 Overview of our system

Our technique automatically processes an individual's sent email archive using text analysis and natural language processing techniques. We identify the names of people, places, events, etc. that appear to be significant in the user's past; we then generate prompts for these terms by identifying sentences that contain the answer term, with the answer blanked out. See Fig. 1 for an illustration. Along with the blank, we ask users to fill in additional judgments about certainty, vividness and recency. The system also captures a large number of features about each question and answer. This technique has the advantage that it can be easily administered online and needs no special equipment. Further, it is completely automatic and can scale to large numbers of users.

We have tested this technique in a study with 35 users, and as we will show later in the paper, our results show validity by corroborating a number of hypotheses we might make about memory. For example, the data demonstrates a pattern of forgetting over time, with accuracy going down with the age of the prompt.

1.2 Contributions

Our major contributions in this paper are the following.

- A system and design for scalable testing of autobiographical memory using email archives. The technique is ecologically valid as it tests memory about entities present in the user's past.
- We describe specific techniques for identifying questions and answers from a user's email archive, which we uncovered with an iterative design methodology.
- We provide results from a study spanning 1,400 questions and answers from 35 users. A preliminary analysis of this dataset shows that it is robust to known hypothesis that we might make. It allows future study of other parameters such as certainty, time judgments, sentimental associations, etc. with respect to memory.

The rest of this paper is organized as follows. After a survey of related work, we describe our system design and study setup. We then describe our study results, the dataset that it generated, and some data analysis results. Next, we discuss possible implications of this research, and avenues for future work. We conclude with some learnings from this project.

2. FEATURES OF AUTOBIOGRAPHICAL MEMORY

Tulving described episodic memory as consisting of stored information about autobiographical events and associated context [17]. An example of an episodic, autobiographical memory might be when one said goodbye to a special friend and the context might consist of the specific location, the weather, or the feelings one experienced at the time. An event typically involves specific people and places and their proper names. There are several conceptual distinctions that memory researchers debate, and one of them involves the degree to which a memory reflects general knowledge separate from any particular event. This form of memory is described as semantic, as opposed to episodic memory. For example if one were asked to complete the following sentence, "I went to Paris, ——", even though the memory was for an event, one could deduce the information based on the knowledge that Paris is in France. Typically episodes involve a cluster of contextual information such as the specific people, places, time, and other information that contribute to a vivid re-experiencing of the event. A typical method capturing the degree to which a memory is episodic is to ask subjects to make judgments about the memory such as "remember" versus "know" [22].

There is also research showing that emotional episodes tend to be well remembered [6]. With the development of sentiment analysis, it is be possible to study the relationship between emotional versus non-emotional text in large datasets. People are also conscious of episodic memories in that they are aware of where they acquired them so that if you ask someone how certain they are that an event happened, there should be an association between how certain they are and how accurately they remember it. A final feature of episodic memory is that it tends to fade and become less vivid or accurate over time.

3. RELATED WORK

3.0.1 Models of autobiographical memory

Most people and places vary in personal relevance over our lifespan and our ability to remember these names also varies. For example, it is easy to remember high school classmates and lecture halls yet over time these names and places are more difficult to recall [21]. One approach to studying proper names included the study of publicly available faces and names such as the TV Test, a measure that tested whether people remembered television shows that were popular for a limited period of time ([14][15]). This ability to identify faces and link them to names and meaningful reasons as to why they were famous varies depending on the time window. People remember famous names that are more recently popular than those popular in more remotely past epochs [11].

Williams and Hollan frame memory as a series of problem-solving techniques and use partial information clues as the main vehicle for understanding memory [21]. By describing memory recall as a process of context, search, verify, they relate closely to our approach of using personalized contexts to trigger searches. Their research tested subjects on high school classmates names, taking a biographical approach to testing for memory. They conclude that biographical information can model the memory process in many ways. Unfortunately, this approach is not easily scalable because it involves manual research into the subjects history, such as their high school information, for designing the test.

Anderson and Milson et al. model memory after the well-studied information retrieval problem in computer science [1]. They make the argument that higher frequency words specific to subjects are better recognized. They develop an equation of need probability, used to simulate likelihood of memory recall, based on personal history and relevance of cues. While they do propose a model for understanding memory, they offer no scalable method of testing.

There is a great interest in finding ways of improving memory and online tools like Lumosity and Fitbrains that help in cognitive and memory training have gained a lot of popularity. These tools are currently not personalized, and our techniques can help enhance understanding of memory for use in such applications.

3.0.2 Memories from Digital Archives

There is a lot of interest in applications of digital archives and life-logging to collect, preserve, refresh and even transfer memories to others[1].

[1]A recent special issue of the journal Human Computer Interaction devoted itself to this theme [18].

Whittaker et al. have proposed a set of general design principles for digital tools that support memory [20]. Pensieve actively solicits input from the user by periodically emailing personal questions and attempts to create a repository of reminiscences [12]. The YouPivot system facilitates searching for contextually associated activities on a desktop computer [4]. Petrelli and Whittaker contrast digital and physical mementos used in family memories; their fieldwork corroborates our thesis that email is frequently one of the digital sources of memories [13]. Crete-Nishihata et al. have studied the impact of creating multimedia biographies from personal digital pictures, documents, music, etc. on patients with Alzheimer's disease or mild cognitive impairment [3]. They find that these biographies can have a profound effect not just on the patients, but also on their caregivers.

In a series of papers, Lamming et al.et al. attempt to collect and use personal digital archives of subjects [8][9][7], focusing on episodic memory and autobiographical context as ways of assisting and evaluating memory. They develop ParcTab, geo-location tokens, and PEPYS, systems which are personal devices subjects can wear to map their movements and to create a digital archive. However, all of the papers reference difficulties with the accuracy of such data, and these systems have no way to identify the relative significance of events in a subject's life.

3.0.3 Email mining systems

Examples of tools for mining and visualizing email archives include Muse [5], Themail [19] and Tiara [10]. However, none of these techniques are used directly for testing memory.

4. SYSTEM DESIGN

In our study, we ask each participant fill-in-the-blank questions, where the term blanked out is a named entity. These questions are generated automatically from the last one year of a user's sent email. Each question prompt consist of a single sentence. See Fig. 1 for an example. This section describes the features of the system, along with their rationale, and how questions and answers are selected. It turns out it is non-trivial to identify good questions and answers. The techniques we describe below reflect our learnings through several iterations of trying to generate answers that are meaningful and prompts that are of high-enough quality such that the answer may be reasonably guessed.

4.1 Data preparation

Before running text analysis on the participant's email messages to identify questions and answers, we have to first fetch and prepare the data. After the user logs in, we fetch the text of messages in the sent message folder over the past year. Any attachments, HTML formatting of text, images etc. are not fetched. Once we have the text of the message, we identify quoted and forwarded parts of the messages and strip them from the message to avoid generating questions from these parts. This identification is done by recognizing fixed templates, and is relatively robust given that our studies are conducted with Yahoo and Gmail accounts, which follow fairly standard formats for reply-quoting and forwarding.

In addition to indexing the email text, we build up a contact address book, performing entity resolution to match different email addresses belonging to the same person.

4.2 Identifying relevant answers

Our system first selects candidate answer terms by identifying frequently used names in the text of the user's sent messages. We decided to focus on names because they tend to be specific and memorable, as opposed to generic words that are often interchangeable and therefore hard to guess precisely. Moreover, names typically have only one form and can be checked accurately via computer. We use the Stanford Named Entity Recognizer [16] to identify names. Answer terms could be one or more words long. The named entities extracted typically encompass a variety of types: people, places, organizations, event names and so on. Focusing on sent messages means that the answer terms we identify have been typed up repeatedly by the owner of the email, and should therefore be meaningful to her; in contrast, incoming email may have skewed distributions of irrelevant entities.

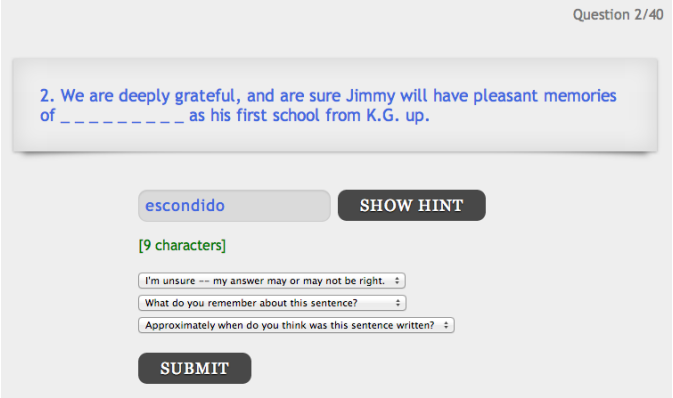

Figure 1: An example of the questions and answers generated by our system. Each question prompt consist of a single sentence, and the answer that has been blanked out is always a named entity. The number of characters is provided to the user. The hint button appears after 15 seconds, and if clicked, reveals the first letter of the answer. The participant is also asked to make 3 judgments about certainty, vividness, and recency.

However, picking just the most frequent terms identified by the entity recognizer as answers for memory testing would not be a good idea. For one thing, the recognizer sometimes misidentifies "names". Over time, we identified a list of words that are commonly misidentified as names by the named entity recognizer, as well as frequently used acronyms such as *FYI, LOL* and *LMAO*. Such words are put in a taboo list so that they are not considered for selection as a possible answer.

Another problem is that, even if a term is frequently used, it could occur in a manner that makes it hard to generate good prompts for it. To solve this problem of some terms not having good prompts for testing, we identify twice the number of terms we are looking for, i.e. to generate N answer terms, we identify a pool of $2N$ candidates. Potential prompts are generated for each of these terms and scored, as described below. Finally, of this pool, the terms that are associated with the top N high-scoring prompts are identified and picked as the answers that will be tested for with the participant.

To address limitations of the NER parser, we also removed names more than 15 characters long, since they generally tended to be long strings that were not actually names, and other names with non-letter characters (these tended to be abbreviated titles as in *Mr. John Ashburn*. These kind of answer terms confuse users because they are unsure about whether titles were to be part of the name. We also remove names from the candidate answers that are prefixes or suffixes of each other. For example, the distinct names *John* and *John Ashburn* will not both be selected as answers in a single session.

Certainty

How sure are you?
1. I have no idea.
2. I'm unsure – my answer may or may not be right.
3. I'm fairly sure
4. I'm certain

Vividness

What do you remember about this sentence?
1. I don't remember anything about it
2. I can infer the answer, but don't recall this context
3. I only recall the general context, not the message
4. I remember this specific message

Recency

Approximately when do you think was this sentence written?
12 options, from the name of the current month, backwards in time.
e.g., September 2013, August 2013, etc.
These are followed by the option "I have no idea".
Information about a wrong answer (asked for wrong answers only)

About this answer...
1. I really should have gotten this correct
2. The answer was on the tip of my tongue
3. My answer is essentially correct
4. This is an insignificant detail that I'm unlikely to have remembered
5. The answer is hard to guess... the clue sentence did not provide enough context

Figure 2: Questions asked in addition to the fill-in-the-blank prompt

4.3 Generating questions

In this section, we discuss a way to generate questions for each of $2N$ candidate answers identified above. Since the answer terms are not restricted to any particular type, a strength of our technique, our prompts have to be able to provide adequate context for any kind of term. We chose a simple and natural strategy to satisfy this requirement, which was to identify sentences containing the answer term, hide the answer, and present the sentence as a fill-in-the-blank question.

However, we still need a way to select a single sentence as the prompt from among all the sentences that contain the answer. Even if an answer is mentioned tens of times, there are different contexts that it occurs in, and each of these might have different qualities as a prompt. For example, a frequent correspondent's name might appear at the beginning of a message in a salutary greeting, but a sentence like "*Hi ——-, Hope you're doing well.*" is not a good prompt for that correspondent's name. Our goal is to weed out questions that might not be appropriate for testing because they provide inadequate context. Note that this issue is somewhat orthogonal to whether the answer itself is intrinsically hard or easy for the user to guess.

We considered many factors in choosing the best prompt for a clue. Some of the features that decide the score of a candidate prompt are:

- Sentiment indicators in the message
- Emoticons and exclamation points in the prompt sentence.
- Length of the sentence
- Sentence number in the message (high sentence numbers due to long messages such as cut and pasted articles generally lead to bad prompts)
- The presence of other names in the sentence.
- Whether the name appears exactly in its form in the sentence, e.g. if the answer term is *John*, a prompt that contains *John Ashburn* may not be appropriate.

The weights for these features in scoring prompts were empirically determined. In the future, the prompts can be weighted towards specific hypotheses under test. Also, note that we did not have our experimental dataset with user feedback on user performance with different prompts. In the future, we can use user provided feedback to learn how to rank prompts, and perhaps even to estimate its level of difficulty.

To identify and tokenize sentences from email, we wrote a custom sentence tokenizer that takes into account emoticons, common abbreviations, etc. While the sentence tokenizer normally ignores newlines, we found it necessary to penalize sentences that span many lines. These tend to be things like bulleted lists, which do not function well as a prompt "sentence". For software engineers, they also tended to be fragments of programming code!

Since question generation can take a few seconds, we generate all the answers and questions up-front, and combine the small delay with the larger time taken to fetch, clean, index and extract named entities from messages. Therefore, there is no delay once the user has started the study.

4.4 Hint mechanism

We note that even prompts for names can be ambiguous and lead to multiple, valid answers, for example, if the user substitutes a nickname or an alternate level of specificity for the correct answer. For example, one might substitute *Bob* for *Robert*, or, in a sentence like *"I met Henner when I visited —— last summer"*, it may be reasonable for someone to enter *"Germany"* for the correct answer *"Munich"*, if one is visiting from another country. We would like to steer users towards the correct form of the answer and allow mechanical checking of answers, with no human involvement. For this purpose, we provide the user with two hints in addition to the prompt. The first is the number of letters in the answer (including the number of letters in each word if the answer consists of multiple words), which is always presented to the user along with the question. The second is an optional hint that provides the first letter of the first word of the answer. This hint appears 15 seconds after the question is presented. This approach to cuing is typical of clinical measures of object naming. It allows a person the chance to generate the name spontaneously. In the case of a "tip of the tongue" phenomenon when a person knows the answer but can not access the particular proper name, the sound of the first letter enables them to recall the word. If the hint button is pressed, this fact is recorded.

Several other hint mechanisms are possible (our pilot users said they sometimes wished they could see the recipients on the message, or its date), but our primary aim with the first-letter hint and providing the number of letters was to help the user disambiguate between several nearly-correct options. We may experiment with different hint mechanisms in the future.

4.5 Additional user judgments

In addition to the answer, we ask the user to provide subjective judgments about how certainty, vividness and recency for each question. At the end, we also ask them to provide some information about wrong answers. We defined the options inductively using our experiences over the pilot study stage; the precise questions asked and the options offered are described in Fig. 2.

The certainty option attempts to gauge how confident the user is about the answers. The vividness option asks the user how he or she arrived at the answer. For example, we observed in our initial testing that users used a variety of strategies to guess the answer. Sometimes, the user remembered the specific sentence being used as the prompt. At other times, the user knew the answer from general context rather than the specific message. Similarly, sometimes it was possible for them to deduce the answer from semantic knowledge, instead of direct memory of an episode. The recency judgment asks the user to guess the approximate month in which the sentence was written. We created the options for wrong answers by observing the common reasons for incorrect answers during pilot testing, described next.

4.6 Pilot testing

Before we launched our formal study, we (the authors) conducted several rounds of testing with our own email archives. After we had the basic system running, we conducted ongoing pilot studies with about 15 individuals as we iterated the design over several weeks. During the iterative process, we would ask a pilot user to run the study, debrief with them on the quality of the test, and implement refinements in response. For example, issues with the processing of data led to sentences not just from users' emails but from their friends' replies. Sometimes, lists of names or items showed up, confusing pilot testers, and leading us to demote non well-formed sentences. Interestingly, some users used the hint simply to verify their answer. To identify this usage, we capture the answer text before the hint is clicked as well as the finally submitted answer. User interface issues were also addressed. Some testers complained about having to count the number of letters in their response to make sure. That led us to add feedback in the interface to indicate when the number of letters was correct (the text describing the number of letters turned green, as shown in Fig. 1 – at other times, it was black).

5. STUDY SETUP

When our system design was fairly robust, we obtained permission from our IRB to recruit online participants. We solicited participation mainly over Craigslist and email lists. Note that we did not solicit participants with memory or cognitive disabilities; our goal in this study was to characterize people with fairly normal memories. The study is administered online over the Internet, and only needs the participant to have a standard web browser.

Participants had to certify that they were above 18, were resident in the United States, and had a majority of their email in English. We required Gmail or Google apps or Yahoo accounts in active use, since these accounts have well-known default Sent folder names. Although our system supports arbitrary folders on any IMAP server, we did not want to allow participants to specify folders to use, in order to ensure uniformity between subjects. For Gmail and Google apps, we had the user authenticate directly with Google with the Oauth protocol for extra reassurance about privacy.

After providing consent, participants initially had to go through a screening step, where we checked to ensure that they had sent at least 20 messages each month for each of the past 12 months [2]. This threshold was decided empirically based on pilot testing. We offered the study only to participants who passed screening. We asked them to work uninterrupted and to not refer to their email, ask anyone for help, or use any other aids during the study.

[2]Throughout this paper, we refer to a 30-day interval as a month, and a 90 day interval as a quarter to ensure uniformity of time windows.

Answer term features

First and last date of usage of term
Answer appears in any address book?
Number of messages/threads with the answer
Monthly histogram of usage

User response features

Hint used?
Hint used only for confirmation?
Milliseconds taken to answer
Number of messages that a wrong answer occurs in
Number of messages in which a wrong answer co-occurs with the correct answer

Prompt sentence features

Length of the sentence (number of characters)
Number of named entities in the sentence
Number of emoticons in the sentence
Sentence number in the message

Prompt message features

Sentiment words in message (Categories tracked: superlative, congratulations, grief, anger, confidential, family, religious, love, vacations, racy, emergency, etc.)
Age of the message (number of days)
Span of thread containing the message (number of days)
Number of names in the message
Number of sentences in the message
Characters in the subject line
Answer part of message recipient name?
Number of recipients

Figure 3: A sample of the parameters tracked for each question in our dataset. A total of 80 features including the judgments described in Fig. 2 are tracked.

	Min	Max	Median	Mean
Number of messages	355	6807	1315	1917
Unique names used	214	2552	959	1018
Number of contacts	53	1295	234	357
Time to completion (sec.)	724	2325	1210	1316
Percentage correct	45	92	80	78

Table 1: Distribution of participant characteristics. Number of messages, names and contacts is computed on original content in sent email. Under time to completion (for the 40 question test), an outlier who took 40,623 seconds is omitted.

"I should have remembered this" *(opt. 1)*	77 (25%)
"Tip of the tongue" *(opt. 2)*	18 (6%)
"Insignificant detail" *(opt. 4)*	102 (33%)
"Not enough context" *(opt. 5)*	109 (36%)
Total	306 (100%)

Table 2: Reasons provided for incorrect answers, across all participants. Option 3 is considered as a correct answer.

We showed the participant an example to make sure they understood the usage of the system before showing them the actual questions. When they had completed all the questions, we showed them the answers the computer had marked as wrong, and asked them to provide more information about these questions.

We compensated participants with a $10 gift card. Email messages were discarded as soon as the user had completed the study, and the result of the study was stored within encrypted files on an encrypted file system.

In all, our dataset (submitted along with this paper) consists of data for about 1,400 questions, with about 80 features per question. Fig 3 lists several of these features. The dataset does not include any personally identifying information nor any text of the actual questions or answers.

5.1 Participant statistics

In all, 186 participants attempted screening, of which only 57 passed. Of these 57, 35 participants actually completed the study. (The rest did not launch the actual study; we did not find anyone leaving the study midway). See Fig. 4 for the age distribution of these 35 participants, which is skewed towards the younger range, with a median age of 26 years and a mean of 29.7 years. Unfortunately, we do not have accurate data on gender as several participants failed to enter this information.

Table 1 provides other statistics about the participants and their email collections. The median time taken to answer the 40 questions was 1,240 seconds (about 20 minutes). This does not account for the time to fetch and index the messages, which typically took about 5 minutes for each participant. The total time taken row excludes a single outlier who completed after 40,623 seconds (over 11 hours, probably due to interruptions). The average time was slightly higher at about 1,316 seconds.

5.2 Reasons for incorrect answers

Table 2 shows what participants entered when asked for more information about incorrect answers.

For answers initially marked wrong by the computer, we accepted the user option 3 "My answer is essentially correct" as a correct answer (this happened for 42 out of 1400 or 3% of questions), therefore it is not shown in Fig. 2. This was because manual analysis of a sample of these cases revealed that they were overwhelmingly simple typing errors (E.g. *Glorai* instead of *Gloria*), or answers at the wrong level of specificity (e.g. *Los Angeles* instead of *Disneyland*). We may include a simple check based on edit distance metrics to identify the former cases automatically in the future.

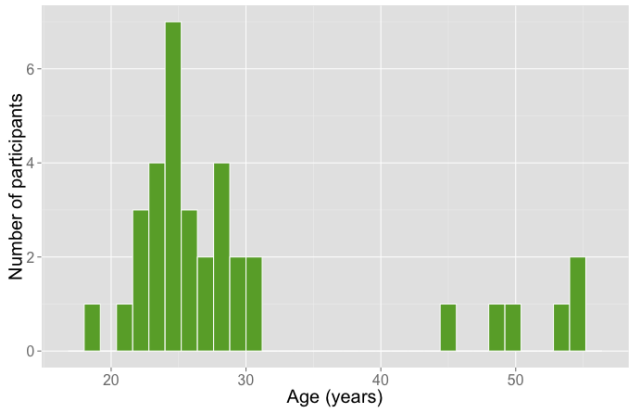

Figure 4: Age distribution of the 35 study participants.

Study participants marked 108 of 1400 questions (7.4%, an average of about 3 questions per user) as not having adequate context. We believe this means our techniques to generate clue sentences are reasonably good. In our early iterations using naive techniques for prompt identification, over 30% of questions would routinely be generic and virtually unanswerable.

6. STUDY RESULTS

In this section, we present some data analysis of our dataset to confirm its validity and explore some initial hypotheses that we had. All 2x2 tests use Fisher's exact test of independence.

6.1 Answer term features

	All	Correct	Incorrect
Frequency of answer (median=8)			
Answer more frequent	699	592 (85%)	107 (15%)
Answer less frequent	701	502 (72%)	199 (28%)
Span of answer (median=176 days)			
High answer span	696	592 (85%)	104 (15%)
Low answer span	704	502 (71%)	202 (29%)
Days since last mention (median=59)			
$>$ median	699	504 (72%)	195 (28%)
$\leq$ median	701	590 (84%)	111 (16%)

Table 3: Correct and incorrect answers by answer term characteristics, using median splits for frequency and span of answer term.

Table 3 shows how participant performance depends on properties of the answer term, divided by a median split. As one might expect, participants performed significantly better on terms that occur more frequently ($p < .001$, odds ratio = 2.19), as well as on terms that span a relatively high period of time (($p < .001$, odds ratio = 2.29). The span is the difference between the first and last day that the term is mentioned. Therefore, the span would be high for terms in ongoing use, and low for terms that are bursty or rare.

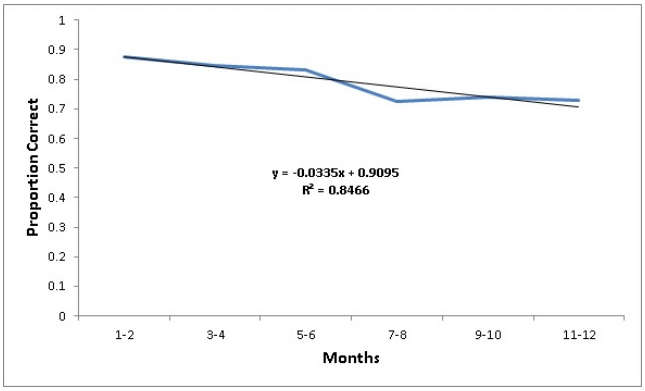

Figure 5: Performance over age of prompt.

(Note that the prompt sentence itself could come from anywhere within the span of the term). Terms that were last mentioned less than the median number of days before the test were also answered more correctly with high significance ($p < .001$, odds ratio = 2.06).

6.2 Memory Decline over Time

Our dataset shows a decline in memory when questions are binned in 2 month intervals. Consistent with expectations, we demonstrated that there was a decline over time in the percent of names correctly generated. A repeated measure ANOVA over the 6 timepoints demonstrated a significant decline ($F(5,150) = 4.04$, $p = .0018$). A linear regression of this decline demonstrated high level of sensitivity (Fig. 5) fitting 82% of the variance.

6.3 Certainty judgments

Participants demonstrated a good awareness for when their answers were correct. There was a significant correlation between level of certainty of correctness and accuracy ($R = .83$, $p < .05$).

6.4 Sentimental Properties

	All	Correct	Incorrect
Love	21	18 (86%)	3 (14%)
Vacation	68	57 (84%)	11 (16%)
Anger	58	48 (83%)	10 (17%)
Exclamation mark	141	114 (81%)	27 (19%)
Memories	153	124 (81%)	29 (19%)
Family	378	301 (80%)	77 (20%)
Emergency	35	28 (80%)	7 (20%)
Congratulations	45	35 (78%)	10 (22%)
Racy	48	37 (77%)	11 (23%)
Superlative	256	192 (75%)	64 (25%)
Confidential	24	18 (75%)	6 (25%)
Emoticon	25	18 (72%)	7 (28%)
Grief	25	18 (72%)	7 (28%)

Table 4: Correct and incorrect answers by emotion indicator.

Fig. 4 shows the number of correct answers by sentimental indicators associated with the prompt. (Note that multiple indicators may be associated with a prompt). The emoticons and exclamations are derived from the actual prompt sentence, while the remaining sentiments are associated with the entire message. The lexicon used for these sentimental categories is borrowed from the Muse project [5] [3]. Of course, some categories like family and vacations do not refer to pure sentiment; however, we included these categories because may be relevant to memory. While, the cell counts are too low to achieve statistical significance, these numbers point to interesting hypotheses. For example, are names associated with love and vacations better remembered than usual, while names associated with grief are somewhat less well remembered? Our dataset raises such questions that can be further studied by directing prompt generation towards specific properties.

6.5 Prompt features

	All	Correct	Incorrect
Prompt length (median=110 characters)			
Longer prompts	688	553 (80%)	135 (20%)
Shorter prompts	712	541 (76%)	171 (24%)
Presence of other names in prompt			
Present	536	446 (83%)	90 (17%)
Absent	864	648 (75%)	216 (25%)
Thread initiated by subject?			
True	599	457 (76%)	142 (24%)
False	801	637 (80%)	164 (20%)
More than one recipient?			
True	319	250 (78%)	69 (22%)
False	1081	844 (78%)	237 (22%)

Table 5: Correct and incorrect answers by prompt characteristics.

Fig. 5 shows how participant performance varies with various features of the prompt sentence, message or thread. The only factor that has a significant influence among these is the presence of names other than the answer in the prompt ($p < .001$, odds ratio = 1.65). This is not surprising since another name in the sentence can provide a valuable clue to the answer. We hypothesized that the length of the prompt, whether the thread was initiated by the user or someone else, and whether the message was addressed to one person or more might make a difference. However, these factors did not significantly impact correctness (with $p = .052$, 0.15, and 0.93, respectively). Our results with these and other features of prompts can help refine our algorithm for prompt selection and may even let us estimate the difficulty of a prompt in advance.

6.6 Recency judgments

For recency judgments, we restrict our analysis to questions that were both answered correctly and with a high degree of certainty (Certainty: Opt. 1 or 2, i.e. certain or fairly sure). Fig. 6 plots error in recency judgments. Since we asked participants to guess the date of the question at a coarse boundary (August 2013, July 2013, etc), we took the 15th of the corresponding month as the date of the guess and subtracted the guessed date with the actual. The scatterplot indicates a lack of systematic bias (M = -10, SD = 55.5, median = -4). We intend to study this phenomenon in greater detail in the future.

6.7 Participant attributes

Statistics of how various attributes of participants affected correctness are listed in Table 6. Performing median splits on age revealed no significant influence on correctness ($p = 0.64$); this is not

[3] Some examples of the terms in different categories:
grief: grief,tragedy,anguish,mourn,condolence,bereave...
family: mom,dad,mother,father,husband,wife,family,kin....
These terms are matched with the text after stemming.

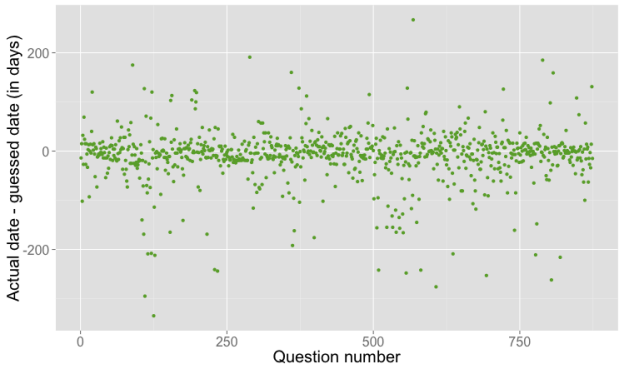

Figure 6: Scatterplot of error in recency judgments for questions answered correctly and with high certainty.

surprising since our participants' age is skewed towards a younger age group.

We also tested for influence of the total number of messages sent in one year and the total number of contacts in the address book, to look for possible effects related to interference. However, perhaps surprisingly, our data shows no significant influence of these factors on correctness (p = 1.0 and 0.43 respectively).

6.8 Qualitative comments

The qualitative comments provided by participants at the end of the study were overwhelmingly positive. 23 of the 35 participants (65%) wrote a comment, and another 8 signed up for the mailing list to be informed about future studies. One of the many positive comments was, *"I enjoyed this study – it was interesting and unique and not too time consuming…"*. One user spoke to the way the test stimulated their memory and said, *"Certainly the most recent messages were easier to remember, but the important events in my life made a difference too.* People used the study to form opinions of their own memory, which ranged from *"I was amazed at how much I DID remember!"* to *"Very interesting study, didn't realize how little I remembered about the emails I send off!"* One person mentioned inadequacy of the provided prompts: *"Cool study but a bit random at times."*.

Study participants echoed a view also expressed by some of our pilot participants: *"I think remembering "when" an email was sent is the most difficult to guess, even when I recall the specific conversation"*.

7. DISCUSSION

This study was able to measure a key property of autobiographical memory, the decrease in accuracy over time. This finding is consistent with previous work on famous proper names such that memory accuracy varies depending the recency of when those people were most popular [11]. Our study was able to capture this decline with great precision in part because of the ability of email to provide detailed time stamps for vast numbers of memory episodes.

The richness of these email archives enables the study of conditions that make memory more and less resilient. For example, just as practice is expected to make memories stronger, these email archives provided meticulous information about how many exchanges were involved in an email interchange so that we were able to demonstrate a relationship to memory accuracy. We provide a rudimentary sentiment analysis for researchers to investigate further whether specific forms of emotional content are more easily remembered (e.g. along the lines of Kensinger & Corkin [6]).

	All	Correct	Incorrect
Count	1400	1094 (78%)	306 (22%)
Hints			
Hint used	454	259 (57%)	195 (43%)
Hint-verify	160	57 (36%)	103 (64%)
Certainty			
Certain	858	830 (97%)	28 (3%)
Certain or fairly sure	1079	1024 (95%)	55 (5%)
Unsure or no idea	321	70 (22%)	251 (78%)
Vividness			
Well-remembered *(opt. 4)*	858	830 (97%)	28 (3%)
Remembered *(opt. 3,4)*	1079	1024 (95%)	55 (5%)
Deduced *(opt. 2)*	108	60 (56%)	48 (44%)
Unremembered *(opt. 1)*	213	10 (5%)	203 (95%)
Participant age (median=26 years)			
$>$ median	640	504 (79%)	136 (21%)
$\leq$ median	760	590 (78%)	170 (22%)
Number of messages (median=1343)			
$>$ median	680	531 (78%)	149 (22%)
$\leq$ median	720	563 (78%)	157 (22%)
Number of contacts (median=234)			
$>$ median	680	525 (77%)	155 (23%)
$\leq$ median	720	569 (79%)	151 (21%)

Table 6: Properties of all, correctly and incorrectly answered questions across all participants.

Another important conceptual breakthrough that results from this paradigm of studying email archives is that the study directly measures peoples' memory for their own lives because the content is drawn directly from their everyday life. In contrast, with standard clinical measures, unless people perform extremely poorly relative to normative samples (e.g. the 2nd percentile of ability), it is difficult to relate test performance to memory ability in everyday life. Another problem with standard lab or clinic based memory measures is that the stimuli are not typically personally relevant. Instead they consist of arbitrary information and developers typically struggle to find content that will be challenging to remember but not so challenging that no one can score correctly on any of the measures. The result is that different measures vary in sensitivity to individual differences in memory ability and there is limited information about features of episodic memory.

Our email tool has important future applications. Because most older adults with dementia resist formal cognitive assessments until there is significant functional disability, there are large web based initiatives to detect the beginnings of memory decline long before it is disabling (thebraininitiative.org). Difficulty with naming is one of the most common complaints of older adults and is disproportionately impaired early in Alzheimer's disease. At present, standard clinical measures of naming are not capable of measuring memory for personally relevant names but this will be possible with our system. It may also be possible to design training programs to strengthen peoples' memories for this personally relevant information. From a scientific standpoint there are several other debates that can be more sensitively addressed. There is controversy around whether different forms of proper names are better remembered, for example, names of people versus places [2]. Generating personally relevant people versus places and capturing the time epoch over which these proper names are active in peoples' emails is easily tested. In sum, the email archives as studied here

present diverse opportunities for clinical and scientific investigation which has previously not been possible.

7.1 Future Work

There are many possible avenues for future work. One line of work is to probe our existing dataset to answer existing research hypotheses. Another is to use the dataset to generate new hypotheses that can lead to further studies focused on specific topics. For example, the association of different types of sentiments with memory may be a rich area to explore using our technique. We may also broaden our context from text-based email to other kinds of digital life-logs, including multimedia. It will be interesting to try and use the results from this study to calibrate and control the expected level of question difficulty.

8. CONCLUSIONS

We have found that it is possible to build ecologically valid tests of autobiographical memory that appear to be accurate, fun and engaging for users. It seems possible that our technique may make an impact on how memory screening tests are conducted, since it is inexpensive and scalable. We hope that our dataset is useful in the future for detailed testing of research hypotheses related to autobiographical memory.

9. ACKNOWLEDGMENTS OMITTED FOR REVIEW

10. REFERENCES

[1] Anderson, J. R., and Milson, R. Human memory: An adaptive perspective. *Psychological Review 96*, 4 (1989), 703.

[2] Cohen, G., and Faulkner, D. Memory for proper names: Age differences in retrieval. *British Journal of Developmental Psychology 4*, 2 (1986), 187–197.

[3] Crete-Nishihata, M., Baecker, R. M., Massimi, M., Ptak, D., Campigotto, R., Kaufman, L. D., Brickman, A. M., Turner, G. R., Steinerman, J. R., and Black, S. E. Reconstructing the Past: Personal Memory Technologies Are Not Just Personal and Not Just for Memory. *Human Computer Interaction 27*, 1-2 (2012).

[4] Hailpern, J., Jitkoff, N., Warr, A., Karahalios, K., Sesek, R., and Shkrob, N. YouPivot: Improving recall with contextual search. In *Proceedings of CHI '11*, ACM (2011).

[5] Hangal, S., Lam, M. S., and Heer, J. MUSE: Reviving Memories Using Email Archives. In *Proceedings of UIST-2011*, ACM (2011).

[6] Kensinger, E., and Corkin, S. Memory enhancement for emotional words: Are emotional words more vividly remembered than neutral words? *Memory & Cognition 31*, 8 (2003), 1169–1180.

[7] Lamming, M., Brown, P., Carter, K., Eldridge, M., Flynn, M., Louie, G., Robinson, P., and Sellen, A. The design of a human memory prosthesis. *The Computer Journal 37*, 3 (1994), 153–163.

[8] Lamming, M., and Flynn, M. Forget-me-not: Intimate computing in support of human memory. In *Proc. FRIEND21, 1994 Int. Symp. on Next Generation Human Interface*, Citeseer (1994), 4.

[9] Lamming, M. G., and Newman, W. M. Activity-based information retrieval technology in support of personal memory. In *IFIP Congress (3)*, vol. 14, Citeseer (1992), 68–81.

[10] Liu, S., Zhou, M. X., Pan, S., Song, Y., Qian, W., Cai, W., and Lian, X. TIARA: Interactive, Topic-Based Visual Text Summarization and Analysis. *ACM Transactions on Intelligent Systems and Technology 3*, 2 (Feb. 2012), 25:1–25:28.

[11] Martins, I., Loureiro, C., Rodrigues, S., Dias, B., and Slade, P. Factors affecting the retrieval of famous names. *Neurological Sciences 31*, 3 (2010), 269–276.

[12] Peesapati, S. T., Schwanda, V., Schultz, J., Lepage, M., Jeong, S., and Cosley, D. Pensieve: supporting everyday reminiscence. In *Proceedings of CHI '10*, ACM (2010).

[13] Petrelli, D., and Whittaker, S. Family memories in the home: contrasting physical and digital mementos. *Personal Ubiquitous Computing 14*, 2 (Feb. 2010), 153–169.

[14] Squire, L., and Fox, M. Assessment of remote memory: Validation of the television test by repeated testing during a 7-year period. *Behavior Research Methods & Instrumentation 12*, 6 (1980), 583–586.

[15] Squire, L. R. Memory Functions as Affected by Electroconvulsive Therapy. *Annals of the New York Academy of Sciences 462*, 1 (1986), 307–314.

[16] Stanford NLP group. The Stanford Named Entity Recognizer.

[17] Tulving, E. *Elements of Episodic Memory*. Oxford University Press, 1983.

[18] van den Hoven, E., Sas, C., and Whittaker, S. Introduction to this Special Issue on Designing for Personal Memories: Past, Present, and Future. *Human Computer Interaction 27*, 1-2 (2012), 1–12.

[19] Viégas, F. B., Golder, S., and Donath, J. Visualizing email content: portraying relationships from conversational histories. In *Proceedings of CHI '06*, ACM (2006).

[20] Whittaker, S., KalnikaitÄŮ, V., Petrelli, D., Sellen, A., Villar, N., Bergman, O., Ilan, B., Clough, P., and Brockmeier, J. Socio-technical Lifelogging: Deriving design principles for a future proof digital past. *HumanâĂŞComputer Interaction 27*, 1-2 (2012), 37–62.

[21] Williams, M. D., and Hollan, J. D. The process of retrieval from very long-term memory. *Cognitive science 5*, 2 (1981), 87–119.

[22] Yonelinas, A. P. The Nature of Recollection and Familiarity: A Review of 30 Years of Research. *Journal of Memory and Language 46*, 3 (2002), 441 – 517.