Enhancing Scientific Collaboration Through Knowledge Base Population and Linking for Meetings

Ning Gao
University of Maryland, College Park
ninggao@umd.edu

Mark Dredze
The Johns Hopkins University
mdredze@cs.jhu.edu

Douglas W. Oard
University of Maryland, College Park
oard@umd.edu

Abstract

Recent research on scientific collaboration shows that distributed interdisciplinary collaborations report comparatively poor outcomes, and the inefficiency of the coordination mechanisms is partially responsible for the problems. To improve information sharing between past collaborators and future team members, or reuse of collaboration records from one project by future researchers, this paper describes systems that automatically construct a knowledge base of the meetings from the calendars of participants, and that then link reference to those meetings found in email messages to the corresponding meeting in the knowledge base. This is work in progress in which experiments with a publicly available corporate email collection with calendar entries show that the knowledge base population function achieves high precision (0.98, meaning that almost all knowledge base entities are actually meetings) and that the accuracy of the linking from email messages to knowledge base entries (0.90) is already quite good.

1. Introduction

Many collaborations in science between distributed and interdisciplinary researchers are inspired by the vision that bringing diverse partners together as a cohesive team can yield more than the sum of its parts. However, studies of actual scientific collaborations sometimes reveal quite different results. For example, a study by Cummings and Kiesler of teams in the NSF Information Technology Research program found that collaborations involving larger numbers of universities and larger numbers of disciplines tended to produce fewer patents and fewer publications. Other studies show that the outcomes of collaborative projects are adversely affected by distance [2] and coordination difficulties [3].

These results have led to increased interest in computational support for coordination and collaboration in distributed and interdisciplinary projects [1, 2, 4]. Despite this interest, a 2005 survey of 71 research projects found that 84% of the teams coordinate using phone or email discussions [4]. That result tends to confirm results reported in 2000 by Olson and Olson showing that the most popular collaboration technologies at the time were telephone, fax, email, audio conferencing, voice mail, and attachments to email. Today we might add videoconferencing services such as Skype, short message apps such as Twitter, shared document editing services such as Google Docs, and shared calendar systems such as Outlook to that list.

Fundamentally, however, the information space of coordination tools remains largely balkanized, with many specialized tools each containing a piece of the puzzle. This poses challenges for new members of a research team, who need to learn to navigate a complex social system in which expertise is distributed in ways that may not be easily discerned. This balkanization also poses even greater challenges for future researchers who might benefit from rich access to the records of completed projects, because many of the support structures available to members of current projects (e.g., disciplinary mentors or local team leaders) will no longer be functioning in those roles.

These considerations have led us to focus on reconstructing links between otherwise disconnected components of a project's information space. In earlier work, we have focused on connecting mentions of people that are found in email messages to the specific people who were being mentioned, the task of entity linking [blinded]. In this paper, we take the next step by broadening our focus to activities, and in particular to meetings. Following the process that we have previously applied to mentions of people, we introduce the task of meeting linking. We first identify as many meetings as we reliably can (in this case, using calendar entries) to construct a knowledge base of meetings, and then we seek to link mentions of meetings in natural language in email to the specific mentioned meeting in the knowledge base. Our work is bottom up in the sense that we are seeking to build technical capabilities that can ultimately be used, both by new members of a project and by future researchers, but our focus at this point is on how well our systems work; we are not yet ready to study how they will actually be used.

The remainder of this paper is organized as follows. The **Background** section reviews some of the relevant prior work. The **Test Collection** section then introduces the collection, which (for practical reasons involving availability and redistribution rights) are drawn from a corporate system rather than from a scientific collaboration. The **System Framework** section gives an overview of the proposed system, and the **System Design** section describes the knowledge base population and linking processes in detail. This is followed by the **Experiments** section, which presents evaluation results for our system. Finally, the **Conclusion and Future Work** section concludes with a discussion of the implications of these results and some thoughts on next steps.

2. Background

Natural language processing can be used in a number of ways to characterize conversational content (e.g., email or recorded teleconferences). Here we focus on knowledge base population and reference linking. Knowledge base population systems can be used to build collection-specific knowledge bases by automatically extracting person [5] or organization entities [6] from a large collection of emails. Entity linking systems [8, 9, 10, 13, 29] have been used to link the named mentions of person, organization and location entities in emails [8, 9] or phone recordings [10, 14] to collection-specific knowledge bases and to general-coverage knowledge bases (e.g., those derived from Wikipedia). When the resulting knowledge base is used to support information access, this can result in improved precision, and sometimes also in improved recall.

Elsayed et al. [5] proposed a method to build a collection-specific person knowledge base for an email collection by treating the email addresses found as senders or recipients in the collection as candidate entities. The person name variants (e.g., first name, last name, nickname) associated with the email addresses are mined from the header, salutation, signature or the email address itself. Gao et al. [6] proposed a method to build a collection-specific organization knowledge base for an email collection by extracting the domain names (e.g., enron.com is the domain name for john.smith@enron.com) from the email addresses in the collection as the set of candidate organization entities. Associated name variants for each organization entity are then mined from four sources: the first returned webpage by posing the domain name as a keyword search to Google, the Wikipedia page that best matches the domain name, organization names found in signature blocks and body of messages sent from that domain.

Entity linking for dissemination-oriented content (e.g., news articles) has been widely studied by researchers for years [16, 24]. However, the task of entity linking for conversational content [8, 9, 10, 25, 26, 27, 28] raises new challenges as the context needed to understand the conversation might not be clearly stated. Also the referent entities (e.g., person, organization) mentioned in the conversations might be absent from the publicly available knowledge bases (e.g., Wikipedia, DBpedia), so that collection-specific knowledge bases are needed. Elsayed et al. [8] proposed a person identify resolution system by using evidence from four sources: the email containing the mention; the path from the email message containing the mention to the root of the discussion (PTR); PTRs containing similar conversational participants, and PTRs containing similar content.

Gao et al. [9] built what is currently the state-of-the-art system for the task of linking person named mentions in email messages to collection-specific person knowledge base by training a supervised learning system with a large set of features. Those features were constructed based on string matching between the mention and the name variants for senders or recipients of the message, social network features, lexical evidence for topical similarity, and tailored features that might suggest the absence of the referent in the knowledge base. Importantly, Gao et al.'s system was also able to recognize mentions that refer to entities absent from the knowledge base (so-called NIL mentions, which indicate that no matching entity exists in the knowledge base). Later, Gao et al. [9] extended the system to link three types of named mentions (i.e., person, organization, location) to multiple knowledge bases (i.e., a general knowledge base built from Wikipedia, and the collection-specific person and organization knowledge).

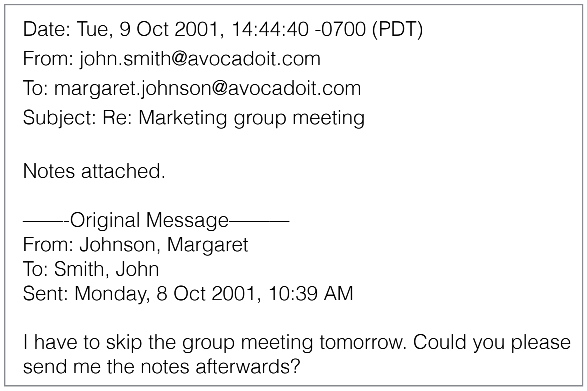
Another related task is the Knowledge Base Acceleration task at the Text Retrieval Conference (TREC) [15]. The knowledge base acceleration task aims to identify documents that contain a mention, given the entity for which a mention is desired. Knowledge base acceleration is intended for filtering a high-volume stream to find documents that could then be mined for attributes of entities to help enrich the knowledge base. Entity and event linking are one step in the knowledge base population pipeline. The Text Analysis Conference (TAC) Event Argument Extraction and Linking shared-task evaluation [11, 12] aims to extract information about entities and the roles they play in events. That task includes a subtask of recognizing mentions of events in dissemination-oriented sources (news articles or discussion forums), for which publicly reported or publicly discussed events (attacks, injury, elections, etc.) are of interest.

Cognitive Assistant that Learns and Organizes (CALO) [17, 18, 19, 20], a project supported by the Defense Advanced Research Projects Agency (DARPA), explored integrating numerous computer-based technologies to assistant users in different levels, including organizing and prioritizing information from different sources (e.g., email, appointments, web pages), mediating human communications by generating meeting transcripts, tracking action item assignments, and detecting roles of participants.

3. Test Collection

For the initial experiments reported in this paper we have used the Avocado email collection [22] available from the Linguistic Data Consortium. The collection contains 614,369 email messages (after de-duplication, for which a standard de-duplicated set is provided with the collection) extracted from 279 email accounts of a defunct information technology company.[1] Most of the accounts are those of Avocado employees, while the remainder of them are shared accounts such as “Marketing Group” or system accounts such as “Conference Room”.

Figure 1 shows a manually constructed example that is similar to email messages found in the collection. Information such as the date sent, senders and recipients (collectively, “participants”), subject, new message content, and quoted text from earlier messages are typically present. There are three types of calendar-like entries within the email accounts: 76,902 appointments (e.g., Communications meeting, system test meeting), 26,980 schedule items (e.g., depart to NY, pick up kids), and 15,473 tasks (e.g., portal update, testing on the hour). In this paper, we focus on the work-related meetings with multiple participants. Most of the “schedule” and “task” entries contain no evidence of discussions between multiple participants. Therefore, we only consider the “appointment” entries when building our collection-specific meeting knowledge base. Figure 2 shows a manually constructed example that is representative of an appointment entry, in this case for a “Marketing Group Meeting”. The owner of the appointment (Margaret Johnson), start time (2001-10-09), recurrence information, and the description of the meeting (located in “text/001/001-000050-AP.txt”) are easily obtained from the XML.



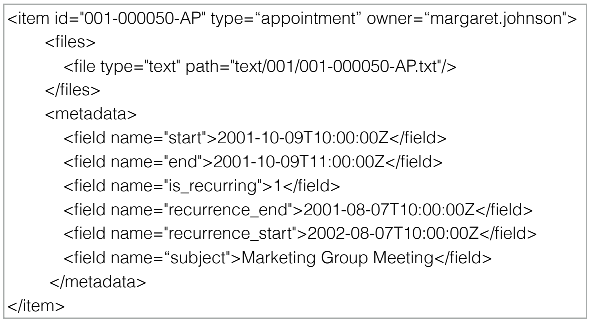
Date: Tue, 9 Oct 2001, 14:44:40 -0700 (PDT)
From: john.smith@avocadoit.com
To: margaret.johnson@avocadoit.com
Subject: Re: Marketing group meeting

Notes attached.

——-Original Message———
From: Johnson, Margaret
To: Smith, John
Sent: Monday, 8 Oct 2001, 10:39 AM

I have to skip the group meeting tomorrow. Could you please send me the notes afterwards?

Figure 1: Email message example.



```
<item id="001-000050-AP" type=“appointment” owner=“margaret.johnson">
    <files>
      <file type="text" path="text/001/001-000050-AP.txt"/>
    </files>
    <metadata>
      <field name="start">2001-10-09T10:00:00Z</field>
      <field name="end">2001-10-09T11:00:00Z</field>
      <field name="is_recurring">1</field>
      <field name="recurrence_end">2001-08-07T10:00:00Z</field>
      <field name="recurrence_start">2002-08-07T10:00:00Z</field>
      <field name=“subject">Marketing Group Meeting</field>
     </metadata>
</item>
```

Figure 2: Appointment entry sample.

There are appointment entries for 226 of the 279 email accounts. Figure 3 shows the number of email messages and number of appointment entries within these email accounts. Each bar represents the number of appointments for an email account, following the scale of y-axis on the right. The line represents the number of email messages for each account following the scale of y-axis on the left. In general, there is no strong correlation between the number of messages and the number of appointment entries (Kendall’s tau [23] is 0.23; where 1 is the strongest positive correlation and 0 indicates no correlation). The email accounts with the most messages are more likely to either be shared accounts (e.g., Marketing Group) or a person who serves as a communication hub (e.g., the president of the company). Similarly, the email accounts with the greatest number of appointment entries are more likely to be shared accounts or meeting coordinators.

Figure 4 shows the number of email messages and appointment entries by year. Again, the line represents the number of email messages following the scale on the left y-axis, and the bars represent the number of appointment entries following the scale on the right y-axis. There is strong correlation (Kendall’s tau of 0.73) between the two distributions. The increasing email activity and the increasing number of meetings between 1994 to 2001 reflects both the

[1] Avocado is a pseudonym, used to refer to the company. As required by the LDC Avocado user agreement, all examples in this paper are manually constructed to be representative of the nature of the content of the collection, but details such as the names of people and the dates and description of events have been changed.

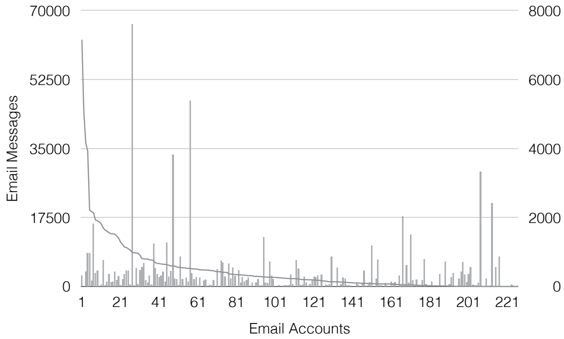


Figure 3: Number of email messages (line) and appointment entries (bars) for each email account, in descending order of the number of email messages.

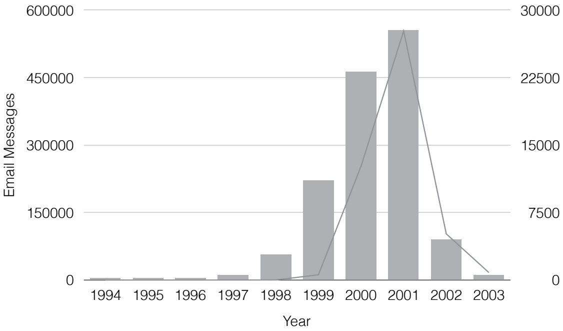


Figure 4: Number of email messages (line) and appointment entries (bars) by year.

growth of the company and the fact that some people retained more older email and calendar entries than did others, while the sharp decrease from 2002 to 2003 might reasonably be interpreted as reflecting changes in the company as it adjusted to new circumstances in the aftermath of the dot com bubble, and then ultimately failed.

4. System Framework

There are five stages in the framework for our system, as shown in Figure 5: collection-specific meeting knowledge base population, query preparation, triaging candidates, feature construction, and prediction. The first step, collection-specific knowledge base population, extracts the appointment entries that are likely to refer to work-related meetings as the meeting entries in the knowledge base. We created guidelines to standardize our definition of a meeting for the experiments reported in this paper: (1) there should be multiple participants in a meeting (e.g., "interview with Greg Kelly" is a meeting, while "Depart at 10:20AM" is not); (2) the owner of the appointment should show intent to go to the meeting (e.g., the owner may go to the "marketing group meeting", but may not for "pizza in the kitchen"); (3) meetings are expected to include some discussion (e.g., calls, video chats, and presentations are considered as meetings, while tasks such as "portal update test" are not); (4) the status indicated in an entry (i.e., Updated, Accepted or Cancelled) does not affect whether it is a meeting (so even cancelled meetings are meetings, since they can be referenced in the text). Appointment entries that meet these criteria were extracted as the candidate meeting entries. Candidate meeting entries that are owned by different accounts are then merged if there is sufficient evidence that they refer to the same meeting. The resulting set of meeting entries constitutes the collection-specific knowledge base. The details of this knowledge base population process are presented in Section 5.1.

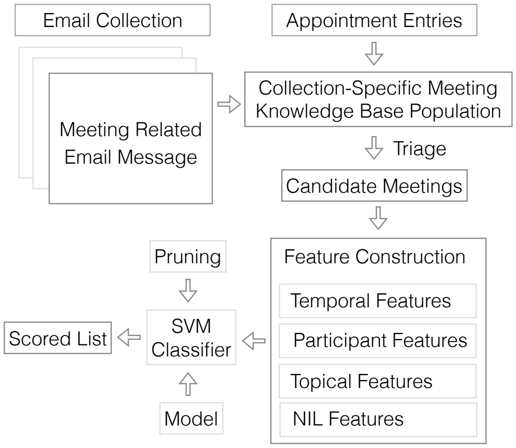


Figure 5: System Framework.

The second step is query preparation. We filter the email collection and select the email messages that contain the string "meet" in either subject or body of the message. Manual annotation of 300 randomly selected email messages by the first author of this paper found that this string match technique achieves a recall 0.98 and a precision 0.79 for identifying messages that contain a mention of a meeting. The false positives include cases when "meet" is referring to a general concept rather than a specific meeting (e.g., no meeting today, meet the requirements). The very few false negatives include cases when the sender of the email messages uses other terms to refer to a meeting (e.g., Call me, let's discuss this tomorrow). According to the manual annotations, 8.9% of the randomly selected messages referred to an existing meeting, while 4.6% of the randomly selected messages contained an invitation to a meeting (e.g., can we meet tomorrow). The remaining 86.5% of the messages were not meeting related.

The third step is candidate triage, in which the goal is to select some (usually) small number of meetings in the knowledge base that could plausibly be the referent of a meeting mention. To do this, indications of the meeting's date are first extracted from the subject and the body of the message. Meeting entries from the knowledge base are then selected as candidates if (1) the meeting is on that date or (if no meeting date indications were found) within some specified time range before or after the date on which the message was sent, and (2) there is at least some participant or topical evidence for the reference. NIL is included as a candidate in every case so that the system has the opportunity to rank NIL along with every other candidate. The details of the triaging step are described in Section 5.2.

For each pair composed of mention of a meeting and a candidate meeting that survives the triage process for that mention, a large set of features are then created in the feature construction stage to calculate the probability that the message is referring to a particular meeting candidate. As explained in Section 5.3, these features are categorized into four groups for presentation purposes. The Support Vector Machine (SVM) regression model nu-SVR from LibSVM [7] is then used with a radial basis function kernel to learn a model that is capable of ranking the candidate meetings for each mention. The top ranked candidate, possibly NIL, is the system's prediction of the meeting to which the mention refers.

5. System Design

We first introduce our knowledge base population technique (Section 5.1), followed by the triage method linking (Section 5.2), and a description for the features used for prediction (Section 5.3).

5.1 Knowledge Base Population

We have built a rule-based system to recognize calendar entries that are likely to be work-related meetings. We first calculate the term frequency of each word appearing in the subjects and descriptions of appointment entries. The 16 most frequently used words (e.g., meet, call) in work-related appointment entries are manually selected as the positive alert list; appointment entries containing one or more words in the positive alert list are candidate meeting entries. Appointment entries with a specific location attribute (e.g., conference room) are also candidate meeting entries. Additionally, appointment entries with known person names in the subject or description (e.g., one on one with John) are considered as candidate meeting entries. To construct the set of known person names we follow the techniques introduced by Elsayed and Oard [5] to first build a collection-specific person knowledge base. The known names are then all known name variants (e.g., first name, last name, nicknames) for every person who has sent or received email in the Avocado email collection.

We also built a negative alert list (e.g., depart, birthday) containing 8 words that are manually selected in a similar manner to recognize appointment entries that do not refer to work-related meetings. Candidate meeting entities containing one or more words in the negative alert list are removed from the candidate set. This process results in a total of 43,499 appointment entries that are recognized as meetings. To evaluate the efficacy of this way of identifying candidate meeting entries, the first author of this paper randomly selected 100 appointment entries and determined whether each entry was a meeting. The system made the same decision as the human annotator on 95 of those 100 cases, for a recall of 97% and a precision of 98%. The same meeting might appear in more than one calendar since every meeting has at least two participants. Any candidate meeting entries that share the same start time, subject and description are therefore merged to produce the final set of meeting entries in the collection-specific meeting knowledge base. A total of 30,449 meeting entries are recognized in this way.

5.2 Linking: Candidate Triage

For each email message containing a detected meeting mention (i.e., each message containing the string "meet"), the candidate triage step of the linking process aims to recognize a small set of meeting entries in the knowledge base that might be the true referent. There are two phases in the triage step. In the first phase, we select the candidates from the knowledge base based on temporal information (e.g., only meeting entries on December 12 can be candidates for email message "feedback for our Dec. 12th meeting"). We first use the Stanford Temporal Tagger [21] to recognize the references to dates (e.g., tomorrow, Thursday, Dec. 12) in the subject field of the email and in the sentences containing the string "meet" in the email body. For example, the sentences "feedback for our Dec. 12th meeting" in a message sent on 2000-12-13, "notes for our Tuesday meeting" in a message sent on 2000-12-10, and "plan for our meeting tomorrow" in a message sent on 2000-12-11 would be recognized and judged as referring to a meeting on 2000-12-12. If a specific date is identified, only the meeting entries on that date are retrieved as the candidates. Otherwise, if any word in the subject field of the email message matched any of the 4 words on a list that we manually created that suggest that the meeting should happen after the message sent date (e.g., agenda, plan) or any of the 5 words on a manually created list that suggest that the meeting occurred before the sent date (e.g., feedback), we take all candidates in a 7-day range on that side of the message. Absent such cues, all the meeting entries within 7 days before or after the sent date of the email message are retrieved as the initial candidate meetings.

In the second triage phase we narrow the list of candidates by searching for the participants or topical contexts matching attributes of each candidate meeting. Candidate meetings with no evidence of being the true referent are removed from the candidate set. We first check the calendars of the email message participants. If the email message is between A and B, then a meeting at which A and B were present could be a potential match. Thus, a meeting is considered as a candidate if it is in two or more than two calendars of the email participants, or if it contains the name of at least one of the participants in the meeting subject or description. Evidence supporting retention could also be found in topical context (e.g., "group meeting with First Tech" could be a candidate for email message "meeting with First Tech"). To check this we extract capitalized words ("Marketing" and "Group" from email message subject or phrase "Marketing Group Meeting"), other than those that contain the string from a 5 manually selected word list (e.g., meet), words indicating time (e.g., Dec.) or status (e.g., Updated), from the subject field of the message and the phrases in the email containing the word "meet" (the phrases are segmented by stop words). A candidate meeting entry is retained if it contains at least one topical term. After this second triage phase, the average number of candidates for each query email message is 11.4 and the median is 6. This two-stage triage process achieves 96% recall on retrieving referenced meetings.

5.3 Linking: Feature Design

Let Q be the email messages in the evaluation set (all of which contain the string "meet"), and M be the collection-specific meeting knowledge base. For each email message $q \in Q$ and meeting $m \in M$, we first identify their extended contexts as $f(P, S, B, T)$, where P represents the participants (sender and recipients) for message q or the owners of meeting m, S is the subject field for q or the meeting subject for m, B is the sentences in the email message body that contain the word "meet" for q or the description of meeting m, and T is the sent date for q or the meeting date for m. Let $\mu_i \in M$ be the set of candidate meetings for query q_i retrieved from the knowledge base M, after triage. We then compute 18 features $\mathfrak{D} = \{\mathcal{D}(q_i, m_j)\}$, where each feature $\mathcal{D}(q_i,\ m_j)$ is expected to have some predictive value for whether a candidate meeting $m_j \in \mu_i$ is the true referent of the meeting mentioned in email message $q_i \in Q$. The features are organized here for presentation purposes into four feature groups by the type of evidence that was used for feature construction.

5.3.1 Temporal Features. This set of 2 features is built based on the temporal information of email message q_i and the candidate meeting m_j. We calculate the unsigned number of days from the email sent date to the meeting date:

$$\mathcal{D}(q_i,\ m_j) \coloneqq |\ T_i - T_j\ |. \qquad (1)$$

There could be multiple dates extracted from the email message by the Stanford Temporal Tagger (e.g., both 2001-10-09 and 2001-10-08 are extracted from the message in figure 1). We therefore use a second feature to calculate the minimum absolute days from the meeting date to any of the extracted dates in the email message.

5.3.2 Participant Features. We build 6 features from the participants in the email message. One feature calculates the number of common participants between email message q_i and candidate meeting m_j:

$$\mathcal{D}(q_i,\ m_j) \coloneqq |\ \{P_i \cap P_j\ \}\ |. \qquad (2)$$

A second feature is Boolean, set to 1 when there are at least two common participants. The other 4 features are based on known name variants (identified as described in section 5.1) for each participant $p\ \in P_i$ in message q_i. Let $N = \{n\}$ be the known name variants for p. We build one feature to calculate number of participants that have any name variant match in the meeting subject

$$\mathcal{D}(q_i,\ m_j) \coloneqq \textstyle\sum_{p\ \in P_i} \mathrm{I}(\left|\{\mathrm{n} \in \mathrm{N}\colon \mathrm{n} \cap S_j \neq \emptyset\}\right| > 0),\ \ (3)$$

where I is the Indicator function. We build another feature to calculate the number of participants that have any name match in the meeting description by substituting B_j for S_j in equation (3)). Finally, we build 2 Boolean features that indicate if there is any name variant match in either the meeting subject or the meeting description.

5.3.3 Topical Features. As mentioned in Section 5.2, terms indicating the topic of the meeting are extracted from the email message in the triage step. We build 4 features based on the term match between email message q_i and candidate m_j. For each message q_i, let $K_i = \{k\}$ be the topic indicative terms.

We build features to calculate the sum of the term frequencies of these terms in the meeting subject S_j:

$$\mathcal{D}(q_i,\ m_j) := \sum_{k \in K_i} \mathrm{TF}(\mathrm{k},\ S_j), \qquad (4)$$

where $\mathrm{TF}(\mathrm{k},\ S_j)$ is the frequency of term k in meeting subject S_j, or the sum of the term frequencies of the topic indicative terms in the meeting description (substituting B_j for S_j in equation (4)). Two additional features are computed by taking the importance of each topic indicative term (as calculated by Inverse Document Frequency in the meeting knowledge base) into consideration (e.g., "Financing" is more informative than "Group" in this context). The subject field feature is computed as:

$$\mathcal{D}(q_i,\ m_j) := \sum_{k \in K_i} \mathrm{TF}(\mathrm{k},\ S_j) \times \mathrm{IDF}(\mathrm{k}), \quad (5)$$

where the Inverse Document Frequency (IDF) of term k is calculated based on the union of the subject and description fields of for each meeting in the knowledge base, defined in equation (6). The description field feature is computed by substituting (the use of B_j for S_j in equation (5)). In general, the more meeting entries the keyword appears in, the less informative it is.

$$\mathrm{IDF}(\mathrm{t}) := \log \frac{1}{|m \in M : k \in \{S_m, B_m\}|} \times |M|, \quad (6)$$

5.3.4 NIL Features. There are 6 features constructed to indicate whether the true referenced meeting might be absent from the knowledge base – the NIL case. We build one feature to indicate if the current candidate is the special NIL candidate that we add to each list (this allows the ranker to learn to treat the NIL candidate differently if that turns out to be helpful). Other features encode: if there are no candidate meetings on the specific meeting date in the query email message q_i; if there is a word (e.g., cancel) indicating the cancellation of the meeting in the message subject S_i; if there is one of those same words indicating the cancellation of the meeting in the topical context B_i; if there is no topic indicative term match in any of the candidates; or if the current candidate meeting m_j is cancelled (with status "Cancelled").

6. Experiments

This section introduces the test collection (Section 6.1) and evaluation measures (Section 6.2). We evaluate the efficacy of linking to known (i.e., Non-NIL) meetings (Section 6.3), separately analyze the utility of each feature group (Section 6.4), and conduct a feature addition study (Section 6.5). Finally, we discuss the linking for NIL cases (Section 6.6).

6.1 Test Collection

Table 1: Statistics on the training and test sets.

	Training	Testing
Known meetings	7,101	7,254
Meeting-related email messages	4,116	7,276
Messages chosen for annotation	617	542
Non-NIL annotations	200	160

To evaluate the efficacy of our meeting linking system we split the email collection and the meeting knowledge base into disjoint training and testing sets. The 226 email accounts with appointment entries are randomly divided into the training and testing sets of equal size. In the training set, we designate as potential "query" email messages those sent on or before 2000-12-31 that contain at least one participant in the training accounts (and the string "meet"). The knowledge base for training is constructed solely from the calendars of the training accounts. In the testing set, the potential query email messages are those sent on or after the date of 2001-01-01 that contain at least one participant in the testing accounts (and the string "meet"). The knowledge base for testing is constructed solely from the calendars of the testing accounts.

Table 1 shows the basic statistics on the training and testing sets. The first author of the paper annotated 617 randomly selected meeting-related email messages (Total annotations) and was able to link 200 messages (Non-NIL annotations) to the meeting entries. For the remaining 417 email messages, the annotator was not able to find the referenced meeting entries either because the true referents are absent from the knowledge base, or because the true referents are difficult for a nonparticipant to find due to the lack of evidence. Three independent annotators were able to link 160 of the 542 randomly selected messages in the testing set to the meeting entries in the knowledge base. The 30% Non-NIL yield on the test set is somewhat lower than the 35% reported for the training set, perhaps because the difficulty of the annotation task may have increased as the company grew, or perhaps because of differences in perspective or ability on the part of the annotators. The 160 Non-NIL annotations in the testing set are used to evaluate the efficacy of our system on linking email messages to the referenced meeting entries. We also evaluate and analyze the system predictions on the NIL links in Section 6.6.

6.2 Evaluation Measures

For each query email message q_i, the set of candidate meetings μ_i will be sorted by the likelihood that they are the true referent according to the SVM regression model. If the true referent is in the candidate set μ_i, let r_i be its rank in the sorted list. If the true referent is not in μ_i, $r_i = +\infty$. We use two metrics to evaluate linking efficacy: Accuracy over all query email messages in Q:

$$\frac{1}{|Q|}\times|\{q_i \in Q: r_i = 1\}|, \qquad (7)$$

and Mean Reciprocal Rank (MRR)

$$\frac{1}{|Q|}\times\sum_{q_i\in Q}\frac{1}{r_i}. \qquad (8)$$

Accuracy shows the fraction of queries for which the top ranked candidate is the same as the human annotation; MRR is a somewhat more forgiving measure that gives partial credit for placing the correct referent lower in the ranked list.

6.3 Linking for Non-NIL

Table 2: Effectiveness measures, Non-NIL queries.

	Accuracy	MRR
SVM	0.899	0.930
Random selection	0.312	0.501

Table 2 shows the efficacy of linking Non-NIL query email messages to the referenced meeting entries. As the random selection results in Table 2 show, this good performance can not be explained solely by imbalances in the data distribution. For example, random selection yields a measured Accuracy of 0.312, which reflects the skewed distribution of triage results. The triage step (Section 5.2) reduces the number of candidates for each query email message from all the meeting entries (7,254) to a median of 6 candidates by taking the temporal, participant and topical information into consideration. After the triage step, 33 of the 160 Non-NIL messages (20.6%) have a single candidate that turns out to be the true referent; these cases account for 0.206 of the 0.312 measured Accuracy of random selection. Our system is able to nearly triple the Accuracy over random selection by using all of our features (Section 5.3). Next, we explicitly analyze the efficacy of each feature group individually (Section 6.4) and in combination (Section 6.5).

6.4 Single Feature Groups

Figure 6 shows the Accuracy for linking the Non-NIL email messages to the referenced meeting entries by using a single group of features. Each bar (Temporal, Participants, Topical, NIL) shows the effect of using all (and only) the features in that group. The Accuracy for random selection and All (using all features) are also shown in Figure 6 for reference. We note that the feature groups used have varying richness along several dimensions, aggregating different numbers of features, with different feature types (binary, integer, of floating point), and different degrees of feature correlation.

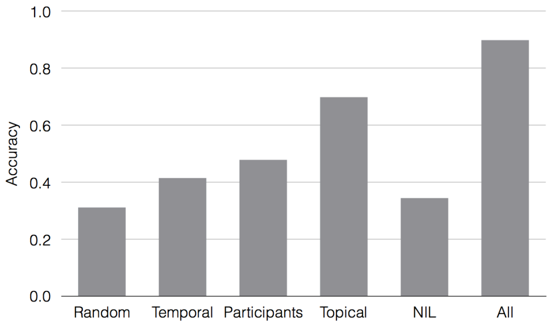


Figure 6: Accuracy for each single feature group.

Topical features are the best single feature group (0.70 Accuracy), and unsurprisingly the features designed for recognizing the absence of the referenced meeting entries (NIL features) result in no improvement when tested on Non-NIL messages. Temporal features are designed to capture the number of days between the email sent date and the meeting date. According to the human annotations, 38% of the meetings mentioned are on the day the email was sent, and 12% of the meeting dates are specified in the email message (e.g., marketing meeting on Dec. 12th). For the reminder of the meetings, email senders are more likely to mention a proximate meeting rather than the one long ago or far in the future. Participant features are designed to search for the names of the email participants in the meeting owners, subjects and descriptions. Within all Non-NIL email messages, 48% have overlap between the meeting owners and the message participants, and 42% contain the names of email participants in the meeting subject or description. Topical features capture the degree of overlap for topic indicative terms (e.g., Marketing) between the email message and meeting entries. On average, less than one keyword (0.69) matches in the true referent, but almost no keywords (0.03) match in the other candidate meeting entries. That sharp difference in distributions is what makes this feature group so useful.

6.5 Feature Group Addition

Figure 6 shows that none of the single feature groups achieves an Accuracy near that of the full set of features. Accuracy thus benefits from the combination of complementary evidence captured by different feature groups. Figure 7 shows the results of

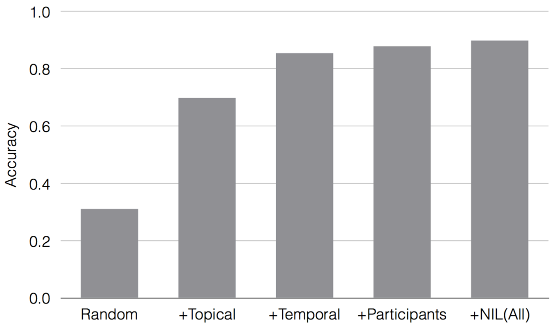


Figure 7: Feature group addition.

cumulatively adding feature groups. From left to right, Random is again the random selection case in which no ranking features are used. We then add the feature group that provides the greatest gain in the Accuracy (Topical, see Section 6.4) yielding an Accuracy of 0.698. Next we try adding each remaining feature set, finding that the combination of Topical and Temporal features achieves the highest Accuracy (0.855). This is close to the result for using all feature groups (0.899). Small improvements result from further adding the most helpful of the two remaining feature sets (Participants) and then from adding NIL features.

6.6 Linking for NIL

In our testing set, our independent human annotators were unable to link 70% of the email messages to a meeting entry in the knowledge base, either because the true referent is absent from the knowledge base or because of insufficient evidence. In other words, these NIL annotations conflate true NILs (meetings that are really missing from the knowledge base) with unresolvable mentions. For example, if the annotator saw an email message from John to Margaret asking "Can we schedule a meeting to discuss the Portal Update?" and there are several meetings in the knowledge base between John and Margaret shortly after that, none of which is called "Portal Update" the annotator may simply not be able to reliably infer which meeting, if any, was being referred to. This problem is reminiscent of the conflation of true NILs with unresolvable mentions in the original set of NIL annotations for person entity linking in email [9]. In that case, just as here (and in contrast to entity linking for dissemination-oriented content such as news), the annotator lacks access to the full context that was available to the email sender and recipients at the time that could have helped them to disambiguate the proper referent.

To simulate the human decisions on NIL links and further analyze the cause of NIL links, we therefore artificially create NIL cases by randomly selecting 10% of the Non-NIL email messages and then removing the true referent for each from the knowledge base. This reduces the Accuracy on Non-NIL query email messages to 0.834 (because we do the same in the training set, thus training on 10% fewer Non-NIL cases) and we can now compare this to the Accuracy we obtain on those 16 (i.e., 10% of 160) artificially created true NIL queries in the testing set, which is 0.438. A manual error analysis shows that there are two dominant explanations for why NIL queries are incorrectly assigned a knowledge base entry: misleading evidence, or prediction with low confidence. For an example of the misleading evidence, consider a message sent on 2001-08-08 regarding "Notes for our Marketing group meeting", for which the true referent is the "Marketing group meeting" on 2001-08-06. After removing the true referent from the knowledge base, the system predicts the referent as the "Marketing group meeting" two days earlier on 2001-08-04. Note that a human annotator might make the same mistake in this situation. For an example of low-confidence prediction, consider an email message sent on 2001-08-08 regarding "Meeting with Greg" for which the a true referent is the "one on one with Greg" on 2001-08-08, but for which the system incorrectly predicts the referent as "Meeting with Greg/Mark/John" on 2001-08-08 after the true referent is removed from the knowledge base. In this case a human annotator might hesitate to annotate this link and if so they would mark the query as NIL. Lacking better candidates, however, our system makes a prediction, albeit with low confidence.

7. Conclusion and Future Work

As one step in our broader interest in building links between presently stovepiped collaboration records, we have developed a system to link mentions of meetings found in email messages to a knowledge base of meeting entries built from calendar entries. Our system works quite well when the mention meeting is present in the knowledge base, although our present implementation is a tad overeager to make a link when none should be made; we should also try other classifier designs. Our present results were obtained on a corporate email collection, and some retuning will likely be needed for the way language is used in particular scientific disciplines when, as a natural next step, we apply our system with the records of a distributed scientific collaboration. In future work we are also interested in integrating other sources (e.g., instant messaging or automatically generated teleconference transcripts).

7. Acknowledgements

This work has been supported in part by NSF Grants 1065250 and 1618695 and by a Mellon Foundation Coherence at Scale Doctoral Fellowship. Opinions, findings, conclusions or recommendations are those of the authors and do not necessarily reflect the views of NSF or the Mellon Foundation.

8. References

[1] Cummings, J. N., & Kiesler, S. (2008). Who Collaborates Successfully? Prior Experience Reduces Collaboration Barriers In CSCW, 437-446.

[2] Olson, G. M., & Olson, J. S. (2000). Distance Matters. Human-Computer Interaction, 15(2), 139-178.

[3] Cummings, J. N., & Kiesler, S. (2007). Coordination Costs and Project Outcomes In Multi-University Collaborations. Research Policy, 36(10), 1620-1634.

[4] Cummings, J. N., & Kiesler, S. (2005). Collaborative Research Across Disciplinary And Organizational Boundaries. Social Studies of Science, 35(5), 703-722.

[5] Elsayed, T., & Oard, D. W. (2006). Modeling Identity in Archival Collections of Email: A Preliminary Study. In Conference on Email and Anti-Spam.

[6] Gao, N., Dredze, M., & Oard, D. W. (2016). Knowledge Base Population for Organization Mentions in Email. In AKBC, 24-28.

[7] Chang, C. C., & Lin, C. J. (2011). LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology, 2(3), 27.

[8] Elsayed, T., Oard, D. W., & Namata, G. (2008). Resolving Personal Names in Email Using Context Expansion. In ACL, 941-949.

[9] Gao, N., Dredze, M., & Oard, D.W. (2017). Person Entity Linking in Email with NIL Detection. In JASIST.

[10] Gao, N., Oard, D. W., & Dredze, M.. (2017). Support for Interactive Identification of Mentioned Entities in Conversational Speech. In SIGIR.

[11] Mitamura, T., Liu, Z., & Hovy, E. (2015). Overview of TAC KBP 2015 Event Nugget Track. In TAC.

[12] Song, Z., Bies, A., Strassel, S.,., et al. (2016). Event Nugget and Event Coreference Annotation. In 4th Workshop on EVENTS.

[13] Dredze, M., McNamee, P., Rao, D., Gerber, A., & Finin, T. (2010). Entity Disambiguation For Knowledge Base Population. In COLING, 277-285.

[14] Benton, A., & Dredze, M. (2015). Entity Linking for Spoken Language. In HLT-NAACL, 225-230.

[15] Frank, J. R., Bauer, S. J., Kleiman-Weiner, M., Roberts, D. A., et al. (2013). Evaluating Stream Filtering for Entity Profile Updates for TREC 2013. In TREC.

[16] McNamee, P., & Dang, H. T. (2009). Overview of the TAC 2009 Knowledge Base Population Track. In TAC..

[17] Tur, G., Stolcke, A., Voss, L., et al. (2008). The CALO Meeting Speech Recognition and Understanding System. In Spoken Language Technology Workshop, 69-72.

[18] Voss, L. L., & Ehlen, P. (2007). The CALO Meeting Assistant. In HLT-NAACL, 17-18.

[19] Ambite, J. L., Chaudhri, V. K., Fikes, R., Jenkins, J., Mishra, S., Muslea, M., & Yang, G. (2006). Design and Implementation of The CALO Query Manager. In IAAI (Vol. 21, No. 2, p. 1751).

[20] Tur, G., Stolcke, A., Voss, L., et al. (2010). The CALO Meeting Assistant System. IEEE Transactions on Audio, Speech, and Language Processing, 18(6), 1601-1611.

[21] Chang, A. X., & Manning, C. D. (2012). SUTime: A library for recognizing and normalizing time expressions. In LREC, 3735-3740.

[22] LDC. (2015). Avocado Research Email Collection. Retrieved from https://catalog.ldc.upenn.edu/LDC2015T03

[23] Kendall, M. G. (1938). A New Measure of Rank Correlation. Biometrika, 30(1/2), 81-93.

[24] Ji, H., Grishman, R., Dang, H. T., Griffitt, K., & Ellis, J. (2010). Overview of the TAC 2010 Knowledge Base Population Track. In Text Analysis Conference.

[25] Liu, X., Li, Y., Wu, H., Zhou, M., Wei, F., & Lu, Y. (2013). Entity Linking for Tweets. In ACL, 1304-1311.

[26] Guo, S., Chang, M. W., & Kiciman, E. (2013). To Link or Not to Link? A Study on End-to-End Tweet Entity Linking. In HLT-NAACL, 1020-1030.

[27] Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., & Bontcheva, K. (2015). Analysis Of Named Entity Recognition And Linking For Tweets. Information Processing & Management, 51(2), 32-49.

[28] Dredze, M., Andrews, N., & DeYoung, J. (2016). Twitter At The Grammys: A Social Media Corpus For Entity Linking And Disambiguation. In 4th Workshop on Natural Language Processing and Social Media, 20-25.

[29] Cucerzan, S. (2007). Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In EMNLP.