ABSTRACT

Title of dissertation: Towards Population of Knowledge Bases from Conversational Sources

Ning Gao, Doctor of Philosophy, 2018

Dissertation directed by: Professor Douglas W. Oard
College of Information Studies

With an increasing amount of data created daily, it is challenging for users to organize and discover information from massive collections of digital content (e.g., text and speech). The population of knowledge bases requires linking information from unstructured sources (e.g., news articles and web pages) to structured external knowledge bases (e.g., Wikipedia), which has the potential to advance information archiving and access, and to support knowledge discovery and reasoning. Because of the complexity of this task, knowledge base population is composed of multiple sub-tasks, including the entity linking task, defined as linking the mention of entities (e.g., persons, organizations, and locations) found in documents to their referents in external knowledge bases and the event task, defined as extracting related information for events that should be entered in the knowledge base.

Most prior work on tasks related to knowledge base population has focused on dissemination-oriented sources written in the third person (e.g., new articles) that benefit from two characteristics: the content is written in formal language and is to some degree self-contextualized, and the entities mentioned (e.g., persons) are likely

to be widely known to the public so that rich information can be found from existing general knowledge bases (e.g., Wikipedia and DBpedia). The work proposed in this thesis focuses on tasks related to knowledge base population for conversational sources written in the first person (e.g., emails and phone recordings), which offers new challenges. One challenge is that most conversations (e.g., 68% of the person names and 53% of the organization names in Enron emails) refer to entities that are known to the conversational participants but not widely known. Thus, existing entity linking techniques relying on general knowledge bases are not appropriate. Another challenge is that some of the shared context between participants in first-person conversations may be implicit and thus challenging to model, increasing the difficulty, even for human annotators, of identifying the true referents.

This thesis focuses on several tasks relating to the population of knowledge bases for conversational content: the population of collection-specific knowledge bases for organization entities and meetings from email collections; the entity linking task that resolves the mention of three types of entities (person, organization, and location) found in both conversational text (emails) and speech (phone recordings) sources to multiple knowledge bases, including a general knowledge base built from Wikipedia and collection-specific knowledge bases; the meeting linking task that links meeting-related email messages to the referenced meeting entries in the collection-specific meeting knowledge base; and speaker identification techniques to improve the entity linking task for phone recordings without known speakers. Following the model-based evaluation paradigm, three collections (namely, Enron emails, Avocado emails, and Enron phone recordings) are used as the representa-

tions of conversational sources, new test collections are created for each task, and experiments are conducted for each task to evaluate the efficacy of the proposed methods and to provide a comparison to existing state-of-the-art systems. This work has implications in the research fields of e-discovery, scientific collaboration, speaker identification, speech retrieval, and privacy protection.

TOWARDS POPULATION OF KNOWLEDGE BASES FROM CONVERSATIONAL SOURCES

by

Ning Gao

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Associate Professor Mark Dredze (The Johns Hopkins University)
Assistant Professor Vanessa Frias-Martinez
Associate Professor Jennifer Golbeck
Associate Professor David A. Kirsch (Dean's representative)
Professor Douglas W. Oard (Chair/Advisor)

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ASR	Automatic Speech Recognition
CALO	Cognitive Assistant that Learns and Organizes
CMU	Carnage Mellon University
CTS	Conversational Telephone Speech
DARPA	Defense Advanced Research Projects Agency
HLTCOE	Human Language Technology Center of Excellence
IDF	Inverse Document Frequency
KB	Knowledge Base
KBA	Knowledge Base Acceleration
KBP	Knowledge Base Population
LDC	Linguistic Data Consortium
LIWC	Linguistic Inquiry & Word Count
LOC	Location
MRR	Mean Reciprocal Rank
NED	Name Entity Disambiguation
NER	Name Entity Recognition
NIL	The referenced entity is absent from all KBs
Non-NIL	The referenced entity is in one of the KBs
OCR	Optical Character Recognition
ORG	Organization
PER	Person
PTR	Path from the query email message to the root of the discussion
SVM	Support Vector Machine
TAC-KBP	Text Analysis Conference Knowledge Base Population
TF	Term Frequency
TREC	Text Retrieval Conference

List of Notations

a	Email address
A	Email addresses associated with an entity
B	Sentences in the email message body that contain the work "meet"
$\mathcal{B}(t, M')$	Term frequency of t in email messages M'
C	Contact list of an entity
c_{e_p,e_t}	Contact frequency between entity e_p and e_t
D	Domain address
$\mathbb{D}$	Feature set
$\mathcal{D}_k$	A feature
$\mathcal{D}(q_i, e)$	Probability that entity e is the referent for query named mention q_i
$\mathcal{D}\left(y_i,\ m\right)\}$	Probability that meeting m is the referent for query q_i
E	Entity search space
$\mathcal{E}_i$	Candidate set for query q_i
E_i	Participants of an email message or telephone conversation
$\hat{E}_i$	Participants of an email message thread
e_o	An organization entity
E_o	Organization entity set
e_p	A person entity
E_p	Person entity set
E_w	Entities from Wikipedia KB
$f(P, E_i, T_i)$	Representation of a query named mention
$f_o(N, D, A)$	Representation of an organization entity
$f_p(N, C, M',)$	Representation of a person entity
$f_w(N, T)$	Representation of a Wikipedia entity
$f(E, L, B, U)$	Representation of a meeting
g	Number of edges of associated with a speaker in the social network
I	Indicator function
k	A topic indicative term
$\mathcal{K}$	Topic indicative terms
K_m	Collection-specific meeting KB
K_o	Collection-specific organization KB
K_p	Collection-specific person KB
K_w	KB built from Wikipedia
L	Subject line
μ	Query email messages
$\hat{\mu}$	Query email thread
m	A meeting entity
M'	Email messages that contain a certain entity
$\mathcal{M}_i$	Candidate meeting set

n	Name variant
N	Observed names for an entity
$\mathfrak{p}$	The probability of an entity being referred by a name variant
P	Type of the named mention
q_i	Query named mention
Q	A set of named mentions
r	The rank of the ground truth in the list
S	An email collection
s_c	The acoustic prediction score of a candidate speaker
s'_c	The re-ranking score of a candidate speaker
s_p	The score of a speaker pair
T	Word vectors
T_i	Vector of words representing the email message content
$\hat{T}_i$	Vector of words representing the email message thread content
U	Email message sent time or the meeting time
W	Number of calls a speaker was detected using each channel
y	Query meeting email message
$\mathcal{Y}$	Query meeting email message set
ϕ	The NIL candidate

Chapter 1: Introduction

The linking of content found in free text to structured knowledge sources is a useful step for information access, archiving, reasoning and discovery. The Text Analysis Conference Knowledge Base Population track (TAC-KBP) introduced the knowledge base population task in 2009 [88] and then divided the complete task into several sub-tasks [28], shown in figure 1.1: *entity discovery* and *entity linking* to extract the mention of entities from unstructured text and link the recognized mentions to the referenced entities in an existing knowledge base; the *event* task to extract information about events so that the information could be entered into a knowledge base; *slot filling* to extract attributes for the entities in the knowledge base; *relation extraction* to extract relationships between the entities; *NIL detection* to cluster all the identical mentions that are referring to entities absent from the knowledge bases, and the *cold start knowledge base population* to populate a knowledge base from scratch without a pre-existing external knowledge base.

Most of the prior work on tasks related to knowledge base population [65, 135] has focused on dissemination-oriented sources (e.g., news articles). Because authors of content intended for broad dissemination must write for a broad audience, it is common practice to write in a self-contextualizing manner. The rise of social

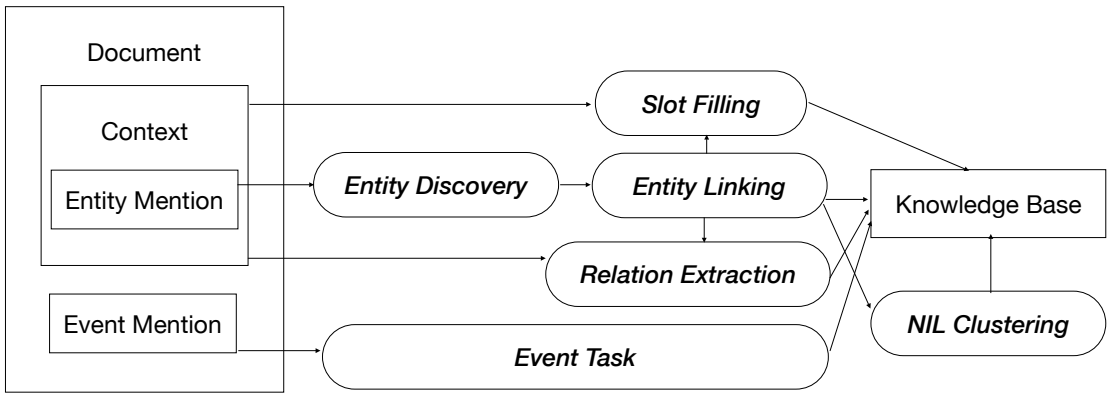

Figure 1.1: Knowledge base population related tasks.

media in recent years has brought fresh attention to what we might call "conversational" sources in which conversational partners interact. This thesis focuses on several tasks that lead to population of knowledge bases for conversational sources, including collection-specific knowledge base population for organizational entities[1] and meetings from email collections; entity linking tasks for three types of mentions (persons, organizations, and locations) in email and phone recording collections; event linking tasks with a particular focus on meeting links for email collections; and the use of speaker identification to improve entity linking for phone recordings.

The recognition and linking of named mentions to real-world entities is the first step in extracting information from unstructured sources. When linking mentions of well-known entities, general-coverage knowledge bases such as those built from Wikipedia are useful search sources for referenced entities. We find in the work described in this thesis (Section 3.2), however, that few people or organizations mentioned in the course of informal interactions exist in such general-coverage knowledge

[1]The types of organizations include corporations, governments, political organizations, international organizations, charities, partnerships, and educational institutions.

bases. This condition prompts renewed interest in constructing collection-specific knowledge bases, which were investigated a decade earlier in other contexts [30]. That approach proved productive, covering approximately 80% of all personal name mentions found in the email collection [31]. The next natural question to explore is whether similar techniques can be used to create collection-specific knowledge bases for organizational entities and events, since the study in this thesis shows that approximately only half (53%) of organizational mentions and none of the work-related meetings mentioned in email collections can be found in Wikipedia. That topic is the focus of Section 3 in this thesis. Table 1.1 concludes the tasks discussed in this thesis.

Given a knowledge base and the recognized name mentions, entity linking, also known as named entity disambiguation (NED), links the name mentions to the referent entities in the KB or returns NIL if the references are absent from the knowledge base. Since their introduction, entity linking studies have explored a variety of data types and settings. Traditionally, many studies have sought to link news articles or web pages to a knowledge base derived from Wikipedia infoboxes [17, 89]. More recently, several studies have considered social media, such as Twitter, as a new source for mentions to be linked to entities [12, 56, 83, 120].

Another thread of work exists on identity resolution in email [24, 31, 99, 137], which is a specialized entity linking task for conversational content. The focus of that work has been on automatically tagging named mentions of a person in the body text of an email message with the email address of that person. If we view the email address inventory as a collection-specific knowledge base, then this task is an

Table 1.1: Tasks studied in this thesis.

Task	**Section(s)**
Collection-specific knowledge base population	3.3, 3.4
Entity linking	4, 6.1
NIL detection	4, 6.1
Event linking	5
Speaker identification	6.2

entity linking task. However, identity resolution research has focused only on person entities, leaving open the opportunity for future work on organizations, locations, and other entity types. Another key problem for identity resolution systems is that all published results on identity resolution in email have been tested only on mentions in which the mentioned entity actually is present in the knowledge base, thus omitting the NIL-detection task. This condition is a severe limitation since the study in this thesis indicates that a substantial number of named mentions in email may refer to entities that are absent from all available knowledge bases. NIL detection is likely to be an important task in many practical applications of entity linking to conversational content. The proposed entity linking system for three types of named mentions (person, organization, and location) with NIL detection is introduced and evaluated on email collections in Section 4 and phone recording collections in Section 6.1.

Entity linking and event linking represent one step in the knowledge base population pipeline. The Text Analysis Conference Event Argument Extraction and Linking shared-task evaluation [100, 124] extracts information about entities and the roles they play in events. That task includes a sub-task of recognizing mentions of events in dissemination-oriented sources (e.g., news articles and discussion

forums), for which publicly reported or publicly discussed events (e.g., attacks, injuries, and elections) are of interest. In this thesis, Section 3.4 explores knowledge base population, and Section 5 studies the linking task for events in email communications with a particular focus on the meeting activities, which is one of the most important coordination and information exchange methods. Meeting activities, as a special type of event, contain a cluster of information, including the participating persons and organizational entities, location of the meeting, subject and description of the meeting, temporal information and other materials. Metadata for meetings (e.g., participants, times, and locations) can be recorded in calendars. Materials, discussions, and meeting notes can be exchanged through emails. Automatically constructing the knowledge base for all meetings in a long-term project and linking all relevant email messages containing related materials tend to improve the efficiency of project coordination, archiving, and semantic searches for meetings 3.4.

The information for writers and readers are typically not included into the entity linking systems for disseminated-oriented sources since the articles are written for a broad audience. However, the experiments in this thesis (Section 4.7 and Section 5.3) show that information about conversational participants plays an important role in distinguishing the referents for entities (particularly for person entities) and meetings mentioned in the conversations. The participants of email conversations can be easily recognized from the headers of email messages. However, the participant information of speech conversations (e.g., phone recordings) is not always available. Speaker identification systems can be applied to automatically recognize the speakers. The efficacy of speaker identification systems can be affected

by additional challenges in practical conversational scenarios (e.g., acoustic conditions such as additive noise or room reverberation), and the specific characteristics of conversations can also provide new opportunities (e.g., in the availability of features that can help to characterize the broader context in which the conversations occurred). Section 6.2 explores the task of coupling acoustic evidence with specific types of side information to improve the performance of the speaker identification task. Section 6.3 explores the use of speaker identification in the entity linking task for conversational speech.

For each of the problems studied, the Enron email collection, Avocado email collection, and Enron phone recording collection are used as the representative repositories of conversational sources. Collection-specific organization and meeting knowledge bases are built through rule-based knowledge base population systems from emails and associated calendars. A supervised machine learning system is built for the task of entity linking, evaluated on email collections and phone recording collection. The efficacy of the proposed linking system is evaluated by accuracy and mean reciprocal rank on randomly selected and human annotated test collections. The cross validation on two email collections shows the stability of the system for Non-NIL mentions. However, for the NIL mentions, high linking accuracy can only be achieved when training and test on the same collection. Although the features are designed based on conversational text collection (i.e., emails), the entity linking system shows comparable high performance for conversation speech collection (i.e., phone recordings) when the speakers are known. However, when the speakers are unknown for the phone recordings, the performance for the person linking

task decreases significantly. In this thesis, speaker identification technique is used to automatically identify the speakers for the phone recordings. Five types of side information are used to improve the speaker identification accuracy. Experimental results show that the improvement achieved by speaker identification task lead to the improvement for entity linking task.

1.1 Research Questions

In this section, the research questions are discussed for the tasks of collection-specific knowledge-base population (Section 1.1.1), entity linking (Section 1.1.2), and meeting linking (Section 1.1.3).

1.1.1 Collection-specific Knowledge-base Population

RQ 1. Can general knowledge bases be used as the linking targets for the mentions of entities in conversational sources?

RQ 2. Are collection-specific knowledge bases needed for the entity linking task?

One initial step towards the knowledge base population challenge is studying the coverage of the existing general knowledge base for the entities mentioned in conversational sources. An exploratory data analysis method is used to study this problem. A test collection is built for linking named mentions of person, organization and location entities from a randomly selected collection of email messages to the general knowledge base Wikipedia. The results show that 68% of the person entities,

47% of the organization entities and 7% of the location entities are absent from Wikipedia. As a conclusion, collection-specific person and organization knowledge bases are needed to resolve the mentions of many entities in conversational sources.

RQ 3. Can collection-specific organization knowledge bases be built from email collections?

RQ 4. How well (in terms of coverage and accuracy) can collection-specific organization knowledge bases be built?

The study of Wikipedia coverage for the referenced entities shows that collection-specific person and organization knowledge bases are needed to resolve the mentions of entities. Collection-specific person knowledge bases for email collections could be built by taking the set of email addresses found as senders or recipients in the collection as candidate entities [30]. To our best knowledge, the construction of collection-specific organization knowledge bases has never been studied until our work in [44]. A rule-based system is built to answer RQ 3, and evaluation is taken to answer RQ 4. This thesis shows that a collection-specific organizational knowledge base with high coverage and accuracy can be built by extracting the domain names found in those email addresses as candidate entities. Both internal targets (email bodies and email signatures) and external targets (Wikipedia and Google search) are used to search for the additional information about the extracted organizational entities.

RQ 5. Can collection-specific meeting knowledge bases be built from calendars?

RQ 6. How well (in terms of precision and recall) can collection-specific meeting knowledge bases be built?

Calendars are potentially useful sources for building a collection-specific meeting knowledge base. RQ 5 discusses whether work-related meeting activities can be recognized from all types of appointments in the calendar (e.g., holidays, dentist appointments, and lunches). RQ 6 discusses whether the necessary information for the meeting activities can be extracted, including the participants, meeting time, location, subject and description. One meeting might be in the calendars of different people, thus raising the third question of whether the meeting entries extracted from different calendars that refer to the same meeting can be merged. A rule-based system is designed to answer the RQ 5, and evaluation is taken to answer RQ 6.

1.1.2 Entity Linking

RQ 7. For the task of linking person named mentions to their referents in knowledge bases, what are useful sources of evidence that could be extracted from email collections, and what are effective ways of using those sources of evidence?

When the referents of the named mentions are in the knowledge bases (the Non-NIL cases), the task of an entity linking system is to recognize the true referent for each mention from all the candidate entities. Prior work in identity resolution (introduced in Section 2.4) complete the task by leveraging evidence extracted from metadata, social and topical context. In this thesis, both new evidence and new ways

of shaping the evidence are introduced to improve the Non-NIL linking efficiency. Hypothesis testing research method is used to anwer this research question. The results show that new features designed in this thesis achieve statistically significant improvement over the state-of-the-art system [31].

> **RQ 8.** For the task of detecting named mentions referring to entities that are absent from all knowledge bases, what are useful sources of evidence, and what are effective ways of using those sources of evidence?

Consider two strong assumptions that generally hold true in TAC-style NIL detection [89]: 1) if many candidate knowledge base entries have names similar to the form of a mention, one of them is likely to be correct; and 2) if no candidates exhibit high similarity between the text of the news story and stored text associated with the candidate entity, the answer is likely NIL. Building features intended to help detect NILs based solely on the number of orthographic (name matching) and topical features can yield high (90%) accuracy for NIL persons [89]. The assumptions that hold for NIL detection in news stories are simply not true for NIL detection in conversational sources. Emails and phone recordings are replete with first-name mentions (e.g., James or Sarah) that might easily match hundreds of candidates yet still be NIL. Thus, relative to the widely studied problem of NIL detection in news stories, NIL detection in conversational sources such as emails can be considerably more challenging. Model-based evaluation is used to answer this research question. The efficacy of the features is evaluated on the test set built on two email collections with annotations for NIL detection.

RQ 9. For the task of linking named mentions of different types of entities to multiple knowledge bases, what are useful sources of evidence, and what are effective ways of using those sources of evidence?

To resolve the named mentions in conversational content, both general knowledge bases and collection-specific knowledge bases are used as linking targets. One challenge is the design of an entity linking system to link all three types of named mentions to all available knowledge bases. Another challenge for conversational content entity linking is that the participants often rely on shared context that may not be explicitly stated in the conversation. This challenge can be addressed by relying in part on social features constructed from the communication graph. Model-based evaluation is used to answer this research question.

RQ 10. Can we make use of side information to improve the speaker identification efficacy for telephone speech?

When the speakers of telephone speech documents are available (e.g., tagged manually), the entity linking system can achieve similar efficacy on the manual transcripts of the speech documents. However, when the ground truth speakers are not provided (e.g., the Enron phone recording collection used in this thesis), a speaker identification system built based on acoustic signals can be used to automatically predict the speakers. Conversations typically occur in a larger context from which we might hope to learn more about the conversation than just what we can obtain from the acoustic signal. Such information might, for example, include associated content (e.g., email among some of the same parties), repeated patterns (e.g., which

phone lines people most often use or whom each person seems to call the most often), or the mention of people at the beginning of the conversation via self-identification of the speakers. The hypothesis testing method is used to answer the question if this side information could be used to improve the efficacy of the speaker identification system built only on acoustic evidence.

RQ 11. Can the efficacy improvement in speaker identification lead to efficacy improvement in entity linking?

For conversational telephone speech, the recognition of speakers plays an important role in the task of person entity linking. However, the speakers are unlabeled for most of the telephone recordings used in this thesis. A speaker identification system is used to automatically predict the speakers for each recording based on acoustic evidences only. Then five types of side information are used to improve the efficacy of the recognition of the speakers. However, one question raises that does the efficacy improvement in speaker identification really lead to the efficacy improvement for person entity linking? The hypothesis testing method is used to answer this research question.

1.1.3 Meeting Linking

RQ 12. For the task of linking meeting-related email messages to the referenced meeting entries in the collection-specific meeting knowledge base, what are the useful sources of evidence, and what are the effective ways of using those sources of evidence?

The study in this thesis shows that meeting-related email messages can be easily recognized with high precision and recall by simply searching for the word "meet" in the subject and body of the messages. With a collection-specific meeting knowledge base available, the question is "What features are most useful for the task of linking the email messages to the referenced meeting entries in the knowledge base?" A supervised machine learning system with a large set of features is the solution studied in this thesis. The efficacy of the proposed system is evaluated on a new evaluation set built for the task.

1.2 Contributions

The contribution of the work in this thesis includes:

1.2.1 Methods

- Development of methods for building collection-specific organization knowledge bases for conversational content (Section 3.3).
- Development of methods for building collection-specific meeting knowledge bases from calendars (Section 3.4).
- Development of a multi-KB structure and methods for entity linking to multiple knowledge bases (Section 4.2).
- Development of methods for linking meeting related email messages to a collection-specific meeting knowledge base (Chapter 5).

- Development of methods for using side information to improve the efficacy of the speaker identification task on conversational telephone speech (Section 6.2.4).

- Development of methods for using speaker identification to improve the efficacy of entity linking task on conversational telephone speech (Section 6.3).

1.2.2 Evaluation

- Introduction of a new measure for the evaluation of the speaker identification task (Section 6.2.2).

1.2.3 Corpora

The released corpora can be found at `http://www.umiacs.umd.edu/~ninggao/publications`.

- Release of the collection-specific organization knowledge bases for Enron and Avocado emails.

- Release of the collection-specific meeting knowledge bases for Avocado emails.

- Release of the new NIL annotations for Elsayed's Enron test collection.

- Release of corpora of human annotations for linking three types of named mentions to multiple knowledge bases for Enron emails, Avocado emails, and Enron Phone Recordings.

- Release of corpora of human annotations for meeting linking task for Avocado emails.

1.3 Applications

The proposed research has a broad applicability including the understanding of speech retrieval results (Section 1.3.1), e-Discovery (Section 1.3.2), scientific collaboration (Section 1.3.3), speaker identification (Section 1.3.4), and privacy protection (Section 1.3.5).

1.3.1 Speech Retrieval

Speech retrieval is a topic of longstanding interest [53]. Early work on speech retrieval focused on formal speech found in news broadcasts, political speeches, and classroom lectures, in part because the accuracy of the automatic speech recognition (ASR) systems used to generate the searched text benefited from the clear articulation, limited vocabulary and formal grammar that is characteristic of formal speech. More recently, satisfactory retrieval results have also been demonstrated for conversational speech [107]. It is becoming increasingly straightforward to create large collections of conversational speech. For example, nearly every teleconferencing service provides such capabilities; certain lifelogging technologies can (when permitted by law) capture conversational speech easily, "talk shows" with debating panelists have become a pervasive element of the media landscape, and video-sharing platforms containing a multitude of types of speech have become ubiquitous. The

question thus arises: "What happens *after* a speech retrieval system has presented conversational speech to the user?" One characteristic of conversational speech is that conversational participants fluidly use references that make sense to them but may be unclear to a person who later encounters that recording, out of context, as the result of a search. For example, the 1,731 freely available telephone recordings used in this thesis were made by Enron employees engaged in regulated energy trading activities. References to "Reliant", "Four Corners", or "Jim" that made sense to conversational partners at the time might be completely opaque to a later searcher who finds a call containing these mentions. The entity linking system introduced in this thesis provides the ability to link specific name references to one or more knowledge bases that can provide additional information about the mentioned entity so that users who are not participants in the conversations can better understand the speech retrieval results.

1.3.2 E-discovery

The task of "e-discovery" refers to the discovery process in litigation or government investigations that addresses the exchange of information in electronic format (often referred to as electronically stored information or ESI). In e-discovery, emails are ubiquitous and constitute more than 50% of the total volume of ESI [16]. In the civil lawsuit brought by the Clinton administration against tobacco companies in 1999 (U.S. vs. Philip Morris), 32 million email records from the White House were made subject to discovery. The rapidly growing volume of such releases makes

it impossible to manually review all the emails in a collection [108]; thus, popular techniques such as keyword searches can help in identifying potential relevant email messages.

For example, in the Enron email collection, to obtain all relevant email messages about James Steffes, former Vice President of Governmental Affairs at Enron, keywords such as "James", "Jim", or "Steffes" might be used as the query keywords. However, one of the problems of retrieving personal mentions by name is that a common first name can easily refer to hundreds of different people in the collection. In the Enron email collection, 760 different people have the same first name "James". As a result, keyword search techniques based on a string match will return all emails (8,240 email messages) about the 760 different "James". The low precision in retrieving relevant emails for person queries necessitates expensive, time-consuming human participation. Using the entity linking system introduced in this thesis, for each of the named mentions (person, organization and location) in the email messages, the referenced entity with associated facts is retrieved from the collection-specific knowledge bases or Wikipedia. Alternatively, if the true referent is absent from all available knowledge bases, then the system will indicate the referent of the named mention as NIL. The problem of searching related email messages for "James Steffes" can be easily rephrased as "return all email messages that contain a named mention referring to the entity James Steffes".

1.3.3 Scientific Collaboration

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 the team's parts. However, studies of actual scientific collaborations sometimes reveal substantially different results. For example, a study by Cummings and Kiesler of teams in the National Science Foundation (NSF) Information Technology Research program found that collaborations involving large numbers of universities and large numbers of disciplines tended to produce fewer patents and fewer publications. Other studies have shown that the outcome of collaborative projects are adversely affected by distance [106] and coordination challenges [20]. These results have led to increased interest in computational support for coordination and collaboration in distributed and interdisciplinary projects [19, 21, 106]. Despite this interest, a 2005 survey of 71 research projects found that 84% of the teams coordinate using phone or email discussions [19]. That result supports the 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 to that list 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.

Fundamentally, however, the information space of coordination tools remains largely balkanized, with many specialized tools each containing a piece of the puz-

zle. This condition 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 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) are no longer functioning in those roles. These considerations point in the direction of reconstructing links between otherwise disconnected components of a project's information space. The meeting linking system introduced in this thesis links the information between email collections and calendars, which has the potential of advancing scientific collaboration by building technical capabilities that can ultimately be used by new members of a project and by future researchers.

1.3.4 Speaker Identification

Understanding conversational speech is a challenging task with many potential applications; examples include providing access to recorded meetings, making sense of the panoply of records that can be generated in lifelogging, and analysis of telephone conversations recorded for regulatory compliance purposes. Some collections that are representative of the use of conversational speech in specific conditions have long been available. Notably, the AMI and AMIDA projects [114] created a corpus of meeting recordings consisting of two types of meetings: a design scenario and naturally occurring meetings in a range of domains. The side information in

those cases includes email messages between the participants. A new collection of Mission Control Center conversations from NASA's Apollo Program is expected to be released soon; in that case, the side information consists of metadata indicating the roles and expected participants on specific intercom circuits, plus thousands of written documents (e.g., technical reports) [104]. Lifelogging is also attracting increasing interest among speech researchers [64, 118, 141], although the first public lifelogging test collection (from NTCIR) focuses on images rather than spoken content [57, 58].

Despite the potential for collections such as these to be used to explore contextual features, the research community as a whole has initially focused their efforts on fully exploiting the acoustic features that are common to all of these applications. In this thesis, experiments were conducted with a conversational telephone speech collection for which five types of side information are available, and have shown that self-trained channel and social network information improve the speaker identification efficacy significantly.

1.3.5 Privacy Protection

The proposed work provides insights and challenges to the field of privacy protection. Conversational text (e.g., emails and short messages) or speech (e.g., telephone recordings) collections are released to the public for different reasons (e.g., research purposes or court orders). To protect the privacy of the people involved in the collection, the name, phone number or other information that might identify

the real-world person are typically redacted. However, even with the identifiable information redacted, the entity linking and collection-specific knowledge base population work introduced in this thesis can sometimes reveal the true identity of a person by integrating the information into the knowledge base. For example, the name of a person entity in the Enron email collection is redacted as X. However, the organization and the office location of X can be extracted from the signature; the title of X could be extracted from the salutations; and the time frame of the extracted information can be found in the email message metadata. Even with the person's name and email address redacted, the identity of X might be easily identified. Thus, a conflict rises between the need to release information and the necessity to protect privacy from automatic systems.

Chapter 2: Background

This chapter reviews the related tasks and studies. Knowledge base population task (reviewed in Section 2.1) is defined as exploring the extraction of information about entities with reference to an external knowledge source (e.g., Wikipedia). Entity linking is one of the most important sub-tasks in Knowledge base population, defined as linking the mentions of entities to their referents in the knowledge bases. Nearly all of the early work in entity linking have focused on third-person reporting (reviewed in Section 2.2), such as news articles or Web pages, where the content is largely self-contextualized. The personal context of the author or reader is less relevant in such settings and thus that has not been a focus of study for entity linking on third-person reporting. Entity linking for twitter (reviewed in Section 2.3) is one step towards the linking task for conversations. The task is more challenging due to the noise, informal language, and implicit context. However, the context of the recipients is still out of the picture. Another related task for the conversational linking is identity resolution (reviewed in Section 2.4), defined as recognizing human identities mentioned in email messages, which could be formed as the person linking task for emails. The linking target for entity linking task has been extended from one single general knowledge base to multiple knowledge bases (reviewed in Section

2.5). The linking source is also extended from text to speech (reviewed in Section 2.6).

2.1 Knowledge Base Population

The Text Analysis Conference (TAC) introduced the Knowledge Base Population (KBP) task in 2009 [88]. Using basic schema for persons, organizations and locations, the entities should be created and populated from unstructured information found in text. Knowledge base population is composed of multiple sub-tasks, many of which have been widely studied for information extraction purposes. Figure 1.1 shows a simple data flow for the sub-tasks in knowledge base population. In each document, the Entity Discovery (i.e., named entity recognition) task recognizes the mentions of entities, while the Entity Linking (i.e., named entity disambiguation) task grounds the recognized mentions to KB entries. The entity mentions that refer to entities absent from the existing knowledge bases are linked to an indicative empty entity called NIL. NIL clustering is the task of clustering all the NIL mentions that are referring to the same entity. The clustered mentions are the sources for new entities in the KB. For each recognized entity in the document, the Slot Filling task is defined as learning the attributes of target entities. The attributes are extracted to enrich the existing KB. Later, two additional sub-tasks were proposed in the KBP track [28], the Cold Start Knowledge Base Population task, defined as building a knowledge base from scratch (i.e., the initial KB in Figure 1.1 is empty) using a given document collection and a predefined KB schema; and the Event task,

defined as extracting information about events such that the information would be suitable as input to a knowledge base.

Despite the various forms and tasks in exploring the population of knowledge bases, there are two goals in building a system, the precision-oriented goal and the recall-oriented goal. Precision-oriented KBP systems [65,66,127,135] aim at building a concrete and precise knowledge base from the documents with high accuracy. Hoffart et al. [65] proposed YAGO2, a knowledge base population system built from only highly reliable resources (Wikipedia, GeoNames and WordNet). The precision of YAGO2 was as high as 0.95. Wolfe et al. [135] proposed a framework of Interactive Knowledge Base Population focusing on extracting information from a small set of topically related documents and constructing Pocket KBs. With an interactive interface, the users were able to visualize and annotate the facts and relations in the knowledge base. The constructed KBs and the models can be improved by adding human-in-the-loop to knowledge base population. The proposed collection-specific knowledge base population for organizations (Section 3.3) and meetings (Section 3.4) in this thesis follows the precision-oriented idea. The methods of extracting entities and the types of attributes are limited to the pre-defined schema. However, the accuracy of the populated information is high enough for practical use.

There is another goal for building KBP systems [4, 37, 78, 87, 110, 123] which is recall oriented. The target is to learn as many facts and relations as possible from a large collection of documents, while the accuracy of the KB might be sacrificed as a result. KELVIN is a cold start knowledge base population system proposed by Mayfield et al. [37, 38, 87, 89–93]. KELVIN was composed of a pipeline of functions

including the discovery of entities, mentions and relations, intra-document coreference resolution, cross-document coreference resolution, inference over the knowledge base, and slot value consolidation. The precision of KELVIN was 0.30 and the recall was 0.47 [93]. Banko et al. [4] proposed an Open Information Extraction (OpenIE) system for the task of Knowledge Base Population. The proposed system extracted a greater diversity of relations that may or may not align to pre-defined relations or to entities that have previously been identified. Later Soderland et al. [123] improved the system by including a human in the loop for manually creating rules to match the automatically extracted relations to the pre-defined rules in TAC KBP track. With 3 hours of human rule creation work, the system achieved precision 0.79 and recall 0.10. Recall-oriented knowledge bases include more types of attributes and relationships, however, the accuracy of the extracted information is impractical to use compared to the precision-oriented knowledge bases such as the ones proposed in this thesis. Assigning confidence to each extracted fact is one possible solution to its practical use in the tasks of question answering and reasoning.

2.2 Dissemination-Oriented Entity Linking

Mihalcea and Csomai [98] defined Wikification as the task of automatically extracting the most important words and phrases in the document, and identifying for each such keyword the appropriate link to a Wikipedia article. A large number of approaches that link named mentions from news articles to Wikipedia entities have been proposed by researchers [12, 15, 17, 27, 35, 36, 60–62, 67, 76, 80, 81, 89, 94,

113, 139, 140].

The TAC KBP track introduced the entity linking task [88] since 2009. The knowledge base was built based on XML data extracted from the October 2008 snapshot of Wikipedia. The entity linking task considers named mentions with certain types: person (PER), organization (ORG) and location (LOC). TAC KBP 2011 [71, 72] further supported NIL clustering and cross-lingual in Entity Linking task. The entity linking system is supposed to extract named mentions from a source collection containing documents in three languages (English, Chinese and Spanish), and link them to an existing general knowledge base (Wikipedia). The system should also cluster the NIL mentions for those referenced entities that are absent from the KB.

Cucerzan [17] proposed an entity linking model by maximizing the agreement between the contextual information, the category tags extracted from Wikipedia and the context of a document. Dredze et al. [27] proposed a flexible entity linking method without depending on the schema of Wikipedia, addressing the problems of robust candidate selection, entity disambiguation, and identifying NIL queries. Hoffart et al. [67] defined three features for identifying referenced entities: popularity prior for entities, context similarity of mentions and entities, and coherence among entities. Latent Dirichlet Allocation (LDA) and its hierarchical variants are natural models in measuring the context similarity between the query document and the candidate entities. This was discussed by the work of Li et al. [82], Kataria et al. [73], Zhang et al. [139], and Bhattacharya and Getoor [7].

McNamee et al. [89] notably brought these ideas together in the context of

the TAC-KBP task to create the Human Language Technology Center of Excellence (HLTCOE) Entity Linker, a two-phase entity linking system. The system first selected candidate KB entities based on triage features. NIL was included in the candidate set and ranked in the same way as any other candidate. In the second phase, the entity linking task was transformed into a supervised learning to rank approach. A set of features (e.g., document similarity, entity classification, popularity, and plausible NIL cues) were computed for each candidate, and the candidates were then ranked by the learned ranking function. The top candidate in the ranked list was returned as the prediction of the referenced entity.

Benton et al. [5] proposed Slinky, an entity linking system with a parallel distributed processing architecture. Slinky allowed cascades with an arbitrary number of stages, processing candidates and queries in parallel. Experiment results showed that Slinky was significantly faster than other non-parallel entity linking systems on large collections.

Work to date on entity linking for dissemination-oriented sources has focused on linking to well known entities, not to individuals that might be known only to a few of the participants. However, when considering interactive communication (e.g., in email), most of the named mentions (e.g., 68% of the person name mentions in Enron [48]) refer to entities who don't have a Wikipedia page.

2.3 Entity Linking for Twitter

There is another research direction focusing on studying user interests through linking named mentions in tweets to the real world entities in Wikipedia [26, 32, 52, 70, 79, 138]. Meij et al. [96] used n-gram to identify potential named mentions, and build features using links within tweets to disambiguate the referred entities in the KB. Genc et al. [52] first mapped the tweets to the most similar Wikipedia pages, and then the distances between Wikipedia pages were used to estimate the distances between tweets.

Liu et al. [83] integrated similarities between mention-entity, entity-entity, and mention-mention to extend the context for tweet linking, and address the problem of name variants. Michelson and Macskassy [97] developed a topic profile for each Twitter user characterizing the topical interests of the users by using the categories containing the most frequently referenced entities. Shen et al. [120] proposed a graph-based model to collectively link the mentions in all tweets posted by one user by reconstructing the topical interest. Guo et al. [56] addressed the tweet linking problem by developing a structural SVM algorithm that jointly optimizes mention detection and entity disambiguation.

The tweet linking task is more challenging than news article collections due to noisy and informal language. However, tweet linking tasks also share the similar characteristics to dissemination-oriented entity linking in terms of self-contextualized and well known entities. For work in tweet linking, we could perhaps model the sender's context, but where the recipient may not be specifically identified. There-

fore, comparing with entity linking for dissemination-oriented content, the implicit context of the participants, both the sender and recipients, provide richer evidence in the conversational content, which offers new opportunities for the entity linking task.

2.4 Identity Resolution in Email

Research on similar problems in email has followed a different path. Klimt and Yang [75] published the CMU Enron email collection. The collection was built from 152 users' email folders, totaling 517,424 messages (without any of the attachments to those messages). The Linguistic Data Consortium (LDC) released the Avocado email collection [105] in 2015. The Avocado collection was built from 279 users' email folders, totaling 614,396 messages after deduplication (with attachments and calendars).

Minkov et al. [99] were the first to pose the identity resolution task, trying an approach based on a structured graphical framework to represent the relationships between identifiable features in an email collection such as content, email addresses, and time. Minkov et al. [99] also built an evaluation collection from Enron email messages using email messages from the folders of two users (Sager and Shapiro) that contained name mentions that corresponded uniquely to the names in the Cc field. To simulate a non-trivial task, the corresponding names were removed from the Cc field for evaluation purposes.

Diehl et al. [24] resolved person named mentions found in the (unquoted)

body text of a subset of Enron by using temporal models of the email traffic. They improved on the Minkov et al. [99] test collection by manually annotating known references for some Non-NIL mentions. This work resulted in a test collection with 78 mentions that are resolved to known Enron email addresses.

Elsayed et al. [31] achieved improved linking accuracy for Non-NIL mentions on the Diehl et al. [24] test collection by using four feature types: (1) presence in the header of the message in which the mention was found, (2) presence in some header in the thread (i.e., reply chain) in which the mention was found, (3) presence in the header of some other threads that contains similar communication participants to the thread in which the mention was found, and (4) presence in the header of some other threads that contains similar content to the thread in which the mention was found. Experimental comparison showed that Elsayed's person identity resolution system achieved better results on Non-NIL mentions (as measured by Mean Reciprocal Rank (MRR) and accuracy) than either Minkov's or Diehl's earlier systems [29].

Elsayed also developed a new, larger test collection using techniques similar to those employed by Diehl et al. [24]. Unlike Diehl et al. [24], however, Elsayed randomly selected mentions from the (quoted or unquoted) body text of randomly selected messages. Mentions that assessors were unable to resolve were labeled as NIL. This allowed Elsayed to address two important limitations of the Diehl et al. [24] test collection: (1) 20% of randomly selected mentions were annotated as NIL, and (2) 16% of the randomly selected mentions resolved to people who did not have an enron.com email address. The Diehl et al. [24] test collection (by design)

includes neither of those categories.

In the only other work on entity linking in email of which we are aware, Xu and Oard proposed an unsupervised ranking model intended for first-stage triage that combines topical similarity features with social network features, again testing their system only on the Non-NIL mentions in Elsayeds collection [137].

If we consider the set of email addresses (together with associated information for each such as known name variants) as a "collection-specific KB" for person entities, then identity resolution is simply a specialized variant of person entity linking. One key difference, however, is that all published results of which we are aware on identity resolution in email before our work in [46] have tested only on entities for which the ground truth entity is in the KB, thus omitting the NIL detection task. Considering the importance of NIL detection in entity linking, the lack of work on NIL detection in this setting limits the practical use of these systems in the task of knowledge base population.

2.5 Entity Linking to Multiple Knowledge Bases

The task of merging KBs for entity linking has been addressed in several studies. Ruiz-Casado et al. [117] and Niemann et al. [103] studied the task of automatically assigning Wikipedia entries to WordNet synsets, which can be considered as simple one-direction merging of two KBs.

Sil et al. [122] proposed an open-database entity linking system that is able to resolve entity mentions detected in text to an arbitrary KB provided in Boyce-

Codd normal form. However, this work focused on distant supervision and domain-adaption, and relied on manually identifying a KB that matches the analyzed documents, without addressing the tasks of detecting domain-specific KBs or maintaining a multi-KB structure.

Demartini et al. [23] used probabilistic reasoning and crowd sourcing techniques for the task of entity linking. Multiple KBs (DBpedia,[1] Freebase,[2] Geonames[3] and NYT[4]) were used as the linking targets. The KBs were simply "combined", and then the candidate entities were triaged by TF-IDF methods.

Pereira [109] proposed resolving the task of entity linking to multiple KBs by using different text and KB features, along with ontology modularization to select entities in the same semantic context, although the detailed structure was not discussed in the paper.

These studies inspire the design of a multiple knowledge base structure in this thesis. The knowledge bases are connected through relationships; for example an organization entity could be linked with a Wikipedia entity if they share the same official URL. Different from previous work, the knowledge bases in this thesis are kept separated rather than merged. Attributes of entities from different knowledge bases are different (e.g., an entity from our collection-specific person KB has the attribute "first name", while an entity from our collection-specific organization KB has the attribute "official website"). Features are also built based on the different attributes of different knowledge bases.

[1] `http://dbpedia.org`
[2] `http://www.freebase.com`
[3] `http://www.geonames.org`
[4] `http://data.nytimes.com`

2.6 Entity Linking for Spoken Language

Benton and Dredze were the first to study entity linking for spoken language [6]. By using Slinky [5], a text-based entity linking system, Benton et al. evaluated the impact of Automatic Speech Recognition (ASR) errors on entity linking using a manually transcribed broadcast news corpus. Experiments showed that the entity linking accuracy drops from 0.77 on manual transcripts to 0.48 on ASR results. The results suggested that spoken language obtained from ASR systems poses more challenges for the task of entity linking: the context can be shorter; transcription errors can distort the context; and named entity recognition tends to have higher error rates. Also feature analysis showed that features built based on phonetic representations of words and expected counts of the lattice for context could improve the accuracy a little bit. Similar to the entity linking task for dissemination oriented content, existing work focuses on entity linking for transcripts of dissemination-oriented speech collections, where the content is self-contextualized and mentioned entities are well known. In this thesis, entity linking for a conversational speech collection is explored for the first time in Chapter 6.

2.7 Other Related Tasks and Systems

Another related task is the Knowledge Base Acceleration (KBA) task at the Text Retrieval Conference (TREC) [1, 8, 22, 25, 40, 54, 84, 85]. In entity linking, the task is to identify a known entity (or return NIL) given a mention. In KBA, by

contrast, the task starts by identifying a document that contains a mention, given the entity for which a mention is desired. KBA is intended for filtering a high-volume stream to automatically recommend edits that can help people to expand the knowledge bases. This focus on streaming content is complementary to our work, in which we presume that the entire collection is available at indexing time, thus enriching the potential for leveraging social network features.

Cognitive Assistant that Learns and Organizes (CALO) [2,131–133], a project supported by the Defense Advanced Research Projects Agency (DARPA), explored integrating numerous computer-based technologies to assist users in different ways, 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. CALO Meeting Assistant (CALO-MA) was an automatic assistant project particularly focusing on organizing the meeting recordings. The client software recorded the audio signals for each meeting participant as well as optional handwriting recorded by digital pens. Automatic Speech Recognition [125, 130] systems were used to transcribe the audio recordings. The transcripts were automatically segmented into sentences with punctuation, capitalization and formatting [18,59,121]. Topic identification and segmentation systems [3,41,68,112] were applied to segment the meeting records into different topics by leveraging lexical features and note-taking behavior. Later the action items and decisions (e.g., task definition, agreement) were extracted [33, 34, 69, 101, 111]. Finally, the meeting summaries were automatically generated [86, 102, 115, 116, 134, 136] by creating a shortened version of the meeting

notes keeping the important points. The CALO project was designed to improve the efficiency of meeting recording and organizing. While this thesis is focused on different tasks regarding the meetings, including constructing a collection-specific meeting knowledge base from the archived calendars, and the meeting linking from email messages to the KB.

Given training speech data from target speakers, speaker identification is the task of determining whether there are target speakers in a test speech segment. The task is divided into text-dependent [77] and text-independent [55]. Text-dependent systems focus on recognizing the speakers from telephone/microphone recorded phone calls, or microphone recorded face-to-face interviews. Text-dependent tasks provide speech collections in scenarios such as different clients pronouncing the same phrase, or each client pronounces his/her own phrase generated by the system. In this thesis, a text-independent system [119] is applied to get system-predicted speakers for the Enron phone recordings. Side information extracted from five different sources are used to improve the speaker identification effectiveness.

2.8 Chapter Summary

In this chapter, related tasks and work were reviewed. Prior work on KBP have been focusing on dissemination-oriented sources, while this thesis focuses on communication-oriented sources (emails, phone recordings, calendars). Collection-specific knowledge bases are constructed following the precision-oriented goal. As a major function in KBP, entity linking for the named mentions (person, organization,

location) are studied. Unlike prior work on identity resolution for emails, the system is able to detect when the true referents are absent from the knowledge bases (i.e., the NIL cases). Also, the problem of entity linking from conversational sources to multiple knowledge bases has never been studied until our work in [49], which will be introduced in this thesis. Event linking has been studied for dissemination-oriented content for years. However, there is few work focusing the events linking on conversational content, which will be discussed in this thesis with a focus on meeting activities. Information of the conversation participants plays an important role in the tasks of entity linking and meeting linking. However, the participants information is unavailable for most of the Enron Phone Recordings used in this thesis. Speaker identification system developed based on solely acoustic signals is applied to get the system estimated speakers. This thesis explores the use of side information to improve speaker identification efficacy, and the use of speaker identification to improve entity linking for conversational speech.

Chapter 3: Knowledge Base Structure

Section 3.1 introduces the two email collections and the phone recording collection used in the experiments. Section 3.2 studies the coverage of the general KB for the named entities in email messages. Section 3.3 describes and evaluates the process of building collection-specific organization KBs from the email collections. Section 3.4 presents the construction of a collection-specific meeting knowledge base.

3.1 Collections

In this section, the Enron and Avocado email collections (3.1.1) used in Chapter 3 and Chapter 4, Avocado calendars (3.1.2) used in Chapter 5, and the Enron Phone Recording collection (3.1.3) used in Chapter 6 are introduced.

3.1.1 Email Collections

Two email collections: Enron and Avocado are used to evaluate the performance of the entity linking system. The first email collection is the CMU version of the Enron email collection, which was built from 152 users' email folders, totaling 248,573 messages after deduplication [75]. The Enron email collection was originally made public by the Federal Energy Regulatory Commission during the investigation

of the Enron Corporate. The CMU version of the collection is originally collected and prepared by the CALO Project (A Cognitive Assistant that Learns and Organizes). The Enron collection has been very widely used in email research, and there is thus some risk of overtuning to a single collection. Replicating experimental results on the newly released Avocado email collection [105] serves to mitigate this risk.

The Avocado collection is built from 279 users' email folders, totaling 614,396 emails after deduplication. Unlike the CMU version of the Enron collection, the Avocado collection includes attachments. The Linguistic Data Consortium (LDC) released the Avocado Research Email Collection in February 2015. The pseudonym "Avocado" refers to a defunct information technology company from which an email collection was created for use by researchers who license that collection from the LDC. Following the techniques in [30], collection-specific knowledge base for the person entities in the collection is first built by using regular expressions to automatically extract associated name-address pairs from the message header and from automatically detected salutations or signatures. For the sender of each email message, associated names are extracted from the salutations and signatures. For each of the email addresses in the header, associated names are extracted by using regular expressions from structures such as "kenneth.lay@enron.com (Ken Lay)". The frequency of each associated email address and name pairs are recorded in the knowledge base as the entity attributes. Then the extracted name-address pairs are merged when there is evidence (e.g., same address) that they are referring to the same person entity. The knowledge base is further cleaned by filtering out the

entities with email accounts formed by numbers only (e.g., 436677@enron.net).

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 3.1: Email message example.

Figure 3.1 shows a manually constructed example that is similar to email messages found in the Avocado[1] and Enron 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.

3.1.2 Calendars

There are three types of calendar-like entries within the Avocado 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 thesis, the work-related meetings with

[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 Avocado collection, but details such as the names of people and the dates and description of events have been changed.

multiple participants are the focus of meeting linking task. Most of the "schedule" and "task" entries contain no evidence of discussions between multiple participants. Therefore, only the "appointment" entries are considered when building the collection-specific meeting knowledge base.

```
<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 3.2: Appointment entry example in Avocado email collection.

Figure 3.2 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. There are appointment entries for 226 of the 279 email accounts.

Figure 3.3 shows the number of email messages and number of appointment entries within these email accounts in Avocado collection. 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

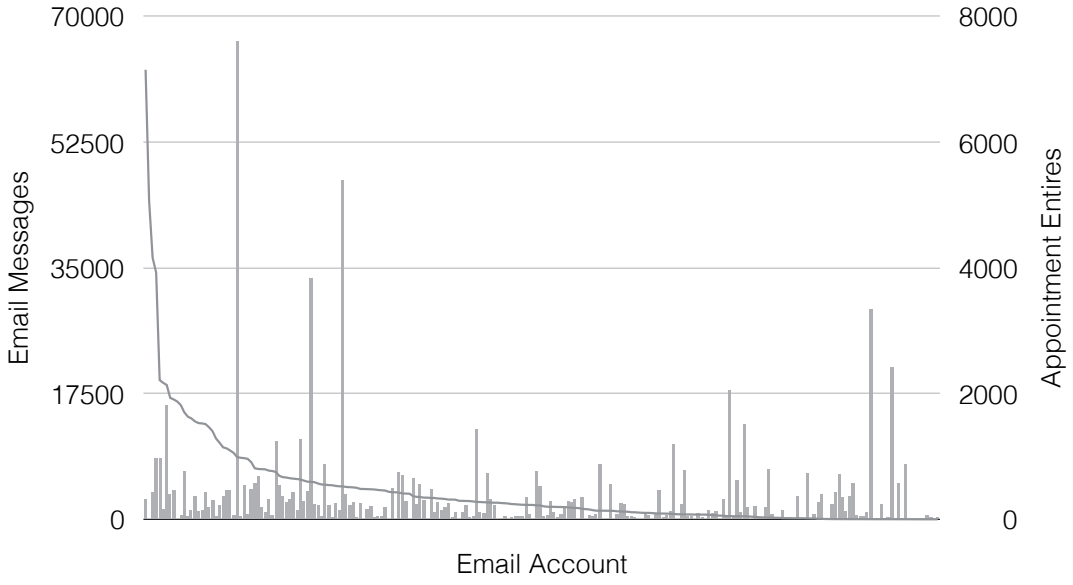

Figure 3.3: Number of email messages (following the scale of y-axis on the left) and appointment entries (following the scale of y-axis on the right) for each email account in Avocado collection.

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 [74] 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 3.4 shows the number of email messages and appointment entries by

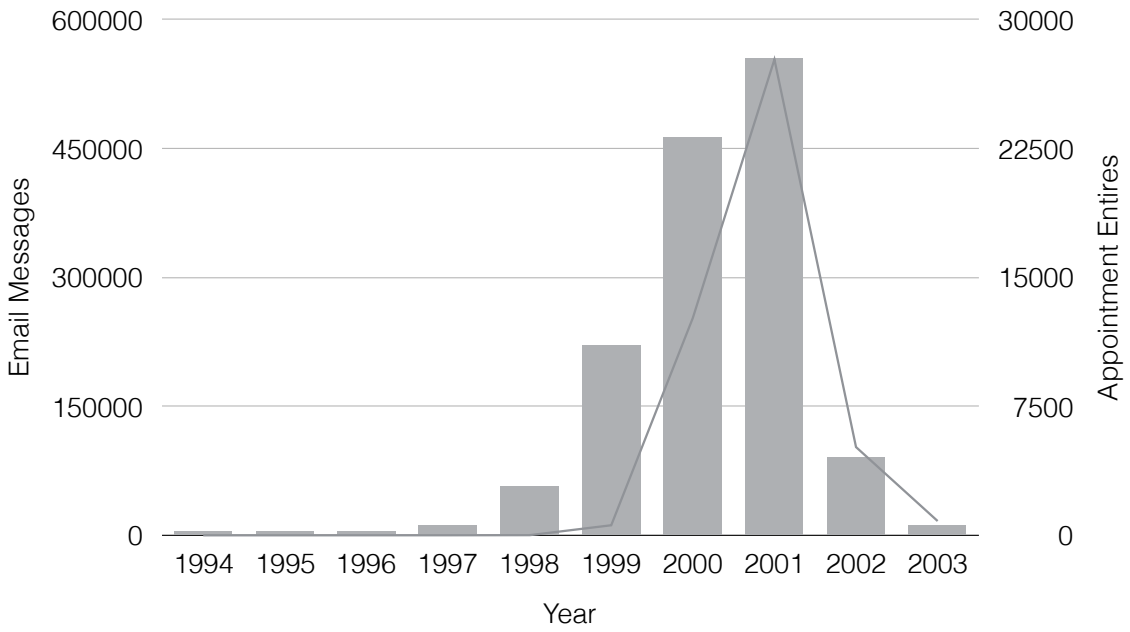

Figure 3.4: Number of email messages (following the scale of y-axis on the left) and appointment entries by year (following the y-axis on the right) in Avocado email collection.

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 activities and the increasing number of meetings between 1994 to 2001 reflects both the growth of the company and the fact that some people retained more emails 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.

3.1.3 Enron Phone Recordings

The Enron Phone Recordings collection contains 1,731 phone recording audio files from Enron traders, which together total 47.8 hours of conversations. Each audio file contains one or more phone calls, and each call includes two or more speakers. These phone recordings were made for regulatory purposes, and were posted to the Internet by the Snohomish County Public Utility District in Oregon pursuant to their use in a lawsuit.[2]

For those recordings that include more than one call, the calls are typically separated by some combination of dial tone, Dual-Tone Multi-Frequency (DTMF) dialing codes, and ring tone. Transcripts were manually prepared for 64 of these recordings for use in court. Those transcripts are available as scanned page images, for which Optical Character Recognition (OCR) yields a low character error rate. We therefore use uncorrected OCR when using these transcripts for content representation (as we do for some of our experiments.) The transcripts include the channel, start time, and duration. Speaker turns in the transcripts are labeled with the name of the speaker when that speaker could be reliably identified by the transcriber; on average there are 1.4 identified speakers per manually transcribed recording (which on average includes 1.5 calls). The document that contains the transcripts also contains a table showing which speakers were either frequently or sometimes observed on each channel. That table had been manually prepared for use in the court case; we transcribed it manually for use in our experiments.

[2] `https://web.archive.org/web/20050206035158/http://www.enrontapes.com/files.html`

The entity linking experiments for conversational telephone speech focus on the manually transcribed 64 files of audio recordings so that the performance on manual transcriptions can be used as a high baseline. The performance on automatic transcriptions (ASR) from the Microsoft Oxford Speech API[3] is also evaluated. ASR systems are highly sensitive to changes in speech content, communication medium, and recording quality. The audio recordings used in this thesis are phone quality conversational audio, which is one of the hardest common tasks for ASR. Therefore, human-readable ASR output is out of expectation, but some correctly transcribed content words can be reasonably expected. In total, these 64 files represent 5.5 hours of conversations. A diarization system run on the 64 recordings shows that the mean duration per speaker turn is 1.99 seconds, and some limited manual diarization confirms that this is a reasonable estimate.

3.2 Coverage of Wikipedia for the Named Mentions in Email

To answer questions "Can general knowledge bases be used as the linking targets for the mentions of entities in conversational sources? Are collection-specific knowledge bases needed for the entity linking task?", a test collection for recognizing and linking named mentions (i.e., PER, ORG and LOC) to Wikipedia is first built based on the deduplicated Enron email collection.[4] Within the CMU Enron email collection [75], there are a substantial number of duplicate email messages (because the same email could, for example, appear in the sender's Sent Mail folder and

[3] `https://www.projectoxford.ai/speech`

[4] This work has been published in Gao, N., Oard, D. W., & Dredze, M. (2014). A test collection for email entity linking. In NIPS Workshop on Automated Knowledge Base Construction [48].

Table 3.1: Statistics for the sampling of named mentions, and the NER accuracy

Type	All Msg	Per Msg $_{\mathbf{All}}$	300 Msg	Per Msg $_{\mathbf{300}}$	Sample	Correct	Accuracy
PER	922,657	3.7	1,262	4.2	200	179	0.895
ORG	1,149,303	4.6	1,879	6.3	200	152	0.760
LOC	492,524	2.0	1,113	3.7	200	193	0.965

the recipient's Inbox folder, and because some users keep multiple copies of email messages, so that the same message might be in the Inbox and also in a folder called "East Oil"). Therefore Elsayed's deduplication process [29] is adopted, in which email messages are considered to be duplicates if they contain exactly the same From, To, Cc, Bcc, Subject, Time, and Body fields. Before deduplication, the date and time of each message are normalized to a standard time zone (Universal Coordinated Time). After deduplication, there are 248,573 email messages in the collection.

To automatically identify named mentions (i.e., omitting nominal and pronominal mentions), the Illinois Named Entity Tagger (INET)[5] is used to recognize three categories of named entity mentions: person (PER), organization (ORG), and location (LOC). In Table 3.1, the column labeled **All Msg** shows the number of named mentions recognized by INET in the whole collection, with the **Per Msg $_{\mathbf{All}}$** column showing the average number of mentions in each email. Then 300 messages are randomly selected that each contained more than 10 words of body text. The **Per Msg $_{\mathbf{300}}$** column shows the average number of mentions in the sampled 300 messages. 20 documents are randomly selected and 135 PER, ORG or LOC mentions are recognized by the author of this thesis. The estimated recall of INET on the sampled documents is 98.5%. Finally, from the detected entities in those 300

[5]INET is available at `http://cogcomp.cs.illinois.edu/page/software_view/4`.

Table 3.2: Wikipedia coverage statistics.

Type	Non-NIL	NIL	NIL %
PER	58	121	68%
ORG	80	72	47%
LOC	180	13	7%

messages (shown in the **300 Msg** column) 200 named mentions (**Sample** size) of each type are randomly selected. Comparing the numbers of **Per Msg $_{\mathbf{All}}$** and **Per Msg $_{\mathbf{300}}$**, the prevalence of named entity mentions is somewhat higher in the sampled collection than in the whole collection because the statistics for the whole collection are reduced by the presence of very short messages of less than 10 words that contain relatively few named mentions.

Six independent annotators then each labeled a different set of 100 of the 600 sampled entity mentions for whether the text span recognized by INET was a correctly delimited named mention of an entity of the corresponding type. For example, in the sentence "I will meet him in Washington DC", the only correct text span would be "Washington DC" (not "Washington" or "in Washington DC"), and "Washington DC" should be classified by INET as a LOC, not a PER. As the **Correct** and **Accuracy** columns in Table 3.1 show, the accuracy for PER and LOC mention detection is comparable to levels typically achieved on news article (90% and 97%, respectively), but detection accuracy for ORG mentions is considerably lower (76%). For the 10% of PER mentions that were missed, informal writing styles in which plausible (but incorrect) names such as "Hope" or "May" can begin a sentence make the task more challenging, particular for systems like INET that are not trained specifically for email.

As expected, there are classes of errors that are also typical in news articles, such as incorrectly segmenting "George Bush" from "George Bush Intercontinental Airport". The few errors on LOC all resulted from typical causes that are also seen in news article (e.g., mis-recognizing "Turkey" as a LOC when, read in context, it is clearly referring to a bird). In the 24% of automatically detected ORG mentions marked by the assessors as incorrect, there are locations mislabeled as ORG (e.g., "QC" in "Montreal, QC") and job positions mislabeled as ORG (e.g., "ITC" in "if you want to be an ITC"). There are also collection-specific names that the assessors simply lacked the knowledge to to judge with confidence (e.g., "RTO" in "standardizing RTO".)

To measure the accuracy of automated entity linking systems, each assessor is asked to provide the correct Wikipedia page for each correctly recognized mention that they assessed, or NIL to indicate if they believed that no such Wikipedia page yet existed. The **Non-NIL** and **NIL** columns in Table 3.2 show the number of entities that the annotator had judged as correct that were or were not in Wikipedia, respectively. A sample of 60 these mentions (20 of each type) was dual annotated by the author, yielding exact agreement (i.e., both designate the same entity or both designate NIL) on 85% of the named mentions. The Cohen's Kappa agreement on whether a mention was NIL or Non-NIL is 0.933.

Table 3.2 shows that only 32% of the PER mentions could be linked to Wikipedia entities. These PER entities include sport stars, politicians, and well known people who worked for Enron such as the former CEO Kenneth Lay. For ORG mentions, about half (53%) of the referenced entities were found in Wikipedia

(e.g., "Justice Department") with the other half annotated as NIL (e.g., "Southward Energy Ltd"). Most (93%) of the LOC mentions could be found in Wikipedia; the relatively few LOC mentions that were resolved by our annotators as NIL included references to specific locations that had not achieved sufficient notoriety for inclusion in Wikipedia (e.g., "1455 Pennsylvania Ave").

Therefore, to resolve the two-thirds of the mentioned person entities and about half of the mentioned organization entities that are not covered by Wikipedia, collection-specific KBs for persons and organizations need to be built. Although Elsayed et al. [30] have produced a collection-specific KB for persons who sent or received messages in the email collections, no comparable collection-specific KB yet exists for organizations.

3.3 Collection Specific Knowledge Base Population for Organization

As shown in Section 3.2, 47% of the mentioned organization entities are absent from the general knowledge base Wikipedia. To resolve these named entities, collection-specific knowledge base needs to be built from the email collection. In this thesis, entities for organizations in the email collection are recognized through extracting domain names in email addresses.[6] Section 3.3.1 describes the process of extracting candidate ORG entities. Section 3.3.2 explains the approach for extracting organization information (e.g., organization name, official website, Wikipedia page) for each entity in the KB.

[6]This work has been published in Gao, N., Dredze, M., & Oard, D. (2016). Knowledge Base Population for Organization Mentions in Email. In Proceedings of the 5th Workshop on Automated Knowledge Base Construction (pp. 24-28) [44].

3.3.1 Extracting Candidate Organization Entities

In the CMU Enron email collection, there are 23,265 unique domain names extracted from the 158,097 unique email addresses in the collection as candidate ORG entities. 22,195 of these domain names have two (e.g., davisbros.com) or three (e.g., dmi.maxinc.com) levels. The remaining 1,070 domain names have between 4 and 6 levels (e.g., dshs.state.texas.us). 39.5% of the unique domain names are associated with at least two different email addresses.

Three steps are applied to regularize the domain names and merge identical ORG entities: (1) All the domain names are lower cased (e.g., ORG entities built from enron.com and ENRON.COM are merged;) (2) Domain name segments within a list of manually recognized words (i.e., main, alert, admin, student, exchange, list) that are not representing affiliations are removed from the domain name; (3) If there are two domain names, and their only difference is ending with *.com* or *.net*, the two associated ORG entities are merged (e.g., entities with domain name enron.com and enron.net are merged), then both .com and .net are dropped from the domain names for each entity. After these simple merging steps, there are 23,008 ORG entities in the collection-specific ORG KB, with each entity contains at least one domain name variant and some number of associated email addresses.

Table 3.3 shows the domain names associated with greatest number of email addresses in the Enron email collection. **Domain Name** represents the domain name extracted from email addresses. **Email Addresses** is the number of associated email addresses. There are two types of email domain names in the col-

Table 3.3: Most Frequently Used Email Address Domain Names.

Top 5 **Organization Domain Names**		Top 5 **Email Service Providers**	
Domain Name	Email Addresses	Domain Name	Emails Addresses
enron	37,687	aol	9,065
haas.berkeley	727	hotmail	6,718
dynegy	633	yahoo	3,919
worldnet	609	msn	1,543
duke-energy	574	earthlink	1043

lection: organization domain names that represent the affiliation of the senders (**Organization Domain Names**), and domain names from large email service providers (**Email Service Providers**). As can be seen, the most frequently used organization domain name is *enron*, followed by the business school of Berkeley *haas.berkeley*, energy company *dynegy*, international courier company *worldnet* and another energy company *duke-energy*. The most frequently used email service providers in the Enron collection are *aol*, *hotmail* and *yahoo*.

3.3.2 Extracting Organization Information

Table 3.4: Extracting Organization Information Through Different Sources.

Source	**Non-NIL Domain Names**	**Email Addresses**	**Non-NIL Accuracy**
Google	68.4%	83%	20/20
Wikipedia	27.6%	64%	15/20
Signature	0.9%	26.3%	20/20
Body	3.4%	29.2%	17/20
Overall	75.1%	87.7%	

Table 3.4 shows the results of using four different sources (Google, Wikipedia, Signature and Body of email message) to extract additional organization attributes. **Non-NIL Domain Names** and **Email Addresses** are the percentages of ORG entities and email addresses that can be associated with an organization name through

certain sources. **Non-NIL Accuracy** shows the accuracy for extracting organization names by manually judging the correctness on a set of randomly selected samples. The details are as follows:

Google. Domain name for each ORG entity is submitted to Google as the search query. If the URL of the top returned webpage contains the domain name, the webpage is considered as the website of the organization. For example, using domain name *bluegate* as search query, the top returned webpage is `http://www.bluegate.com/` with page title *BLUEGATE - Medical Grade Network*. Both the URL and title of the matched webpage are stored as additional attributes for the ORG entities. Corresponding webpages are found for 68.4% of the ORG entities associating with 83% of the email addresses in Enron email collection. If considering only the domain names associated with at least two different email addresses, Google search is able to retrieve websites for 71.2% of the domain names covering 84.6% of the email addresses.

To measure the reliability of the Google source, 20 ORG entities with Non-NIL Google returned webpages are sampled randomly and evaluated manually by the author of this thesis, and all of the webpages are judged as the official organization websites. 20 ORG entities with none (NIL) match between domain names and Google returned webpages are also judged manually, in which 10 of them have changed the organization and website name (e.g., domain name *houston.rr* is changed to domain name *comcast*). For the rest of the 10 NIL ORG entities, no corresponding organization websites are found through Google search.

Wikipedia. For each Wikipedia entity, the URL from *External Links* and *Website* in Infobox are extracted and compared to the domain names for ORG entities in the collection-specific KB. The Wikipedia entity with longest domain name segment match is returned as the additional Wikipedia link for the ORG entities. For example, the *website* (`www.haas.berkeley.edu`) of the Wikipedia entity *Haas School of Business* has the longest domain name segment match with the ORG entity with domain name (haas.berkeley), therefore, the title and website of Wikipedia entity are attached to the corresponding ORG entity in the KB. Through Wikipedia URL match, there are 27.6% of the ORG entities attached with additional Wikipedia entities covering 64% of the email addresses. If considering only the domain names associating with two or more than two different email addresses, 32% of the ORG entities are linked with Wikipedia entities covering 68% of the email addresses.

Manual judgments by the author on 20 randomly selected Non-NIL ORG entities show that 15 (75%) of the entities are matched to the correct Wikipedia entities. When there is more than one segment in the domain name of an ORG entity, it usually represents the hierarchy of the organization (e.g., *store.yahoo* represents *Yahoo Store* in *Yahoo!*). When the Wikipedia entity with the longest domain name segment match is only a partial match, there will be mis-alignment (the 5 errors in the sampled evaluation set) between the ORG in the KB and the organization in Wikipedia. For example, the Wikipedia entity with longest domain name segment match is *Yahoo!* (with *Website* `www.yahoo.com`) for domain name *store.yahoo*, which is a mis-alignment for the organizations in the KB and Wikipedia.

Signature. Signature often contains the affiliation of the sender, so the organization information of the sender's email address domain name can sometimes be extracted from the signatures. Carvalho and Cohen [11] proposed to detect signature blocks in email messages by using a supervised machine learning method. Each email message is represented as a sequence of lines, and each line is represented as a set of features (e.g., line contains URL pattern, line contains phone number pattern, the number of leading tabs equals 2). With the Carvalho and Cohen system, the presence of a signature block in a message can be detected with accuracy 97%, and the signature block lines in a message can be detected with accuracy 99%. The approach proposed in [11] is applied to detect the signatures in email messages.

By using the signature detection results, phrases with capital initials in the signature are recognized as potential organization names if there is a 5-gram string match between the domain name of the sender's email address and the phrase. For example, *Harvard* and *Harvard Business School Publishing* are both valid organization names for domain name *hbsp.harvard*. The frequency of the observed organization name / domain name pairs are also stored for each ORG entity.

By extracting information from the signatures, 0.9% of the ORG entities (covering 26.3% of the mail addresses) are attached with organization names. If considering only the domain names associated with two or more than two different email addresses, organization names can be extracted for 1.3% of the ORG entities covering 28.8% of the email addresses. Manual judgments by the author on 20 randomly sampled Non-NIL ORG entities show that all (20) of the extracted organization names are valid.

Body. Similar to the approach in **Signature**, potential organization information can also be extracted from the body of the email messages. By using the source of email message body, 3.4% of the ORG entities are attached with additional organization names covering 29.2% of the email addresses. If considering only the domain names associated with at least two email addresses, organization information can be extracted for 5.1% of the ORG entities covering 31.8% of the email addresses. Manual judgments by the author on 20 randomly sampled Non-NIL ORG entities show that 13 of the the extracted organization names contain valid information.

Overall. Considering the union of the four sources (Google, Wikipedia, Signature and Body), there are 75.1% of the ORG entities attached with additional organization information covering 87.7% of the email addresses. If considering only the domain names associated with two or more than two different email addresses, organization information (organization names, URLs, Wikipedia page) can be extracted for 77.8% of the ORG entities covering 89% of the email addresses.

3.4 Collection-Specific Knowledge Base for Meetings

In this thesis, the task of linking meeting-related email messages to a collection-specific meeting knowledge base is studied. To complete the task, a collection-specific meeting knowledge base is first built from the calendars. The study is conducted on the Avocado collection.

Manual guidelines are created to standardize the definition of a meeting for the experiments reported in this thesis: (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 (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.

A rule-based system is built to recognize calendar entries that are likely to be work-related meetings. By observing the calendar entries, the author determined that only the calendar entries between the year 1998 and 2002 are to be considered as valid meeting entries. The term frequency is first calculated for each word appearing in the subjects and descriptions of appointment entries. The 16 most frequently used words (meet, call, discuss, presentation, talk, training, plan, review, interview, overview, demo, market, mtg, accepted, introduction, occasion) 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 the techniques introduced by Elsayed and Oard [30] are applied 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.

A negative alert list (depart, birthday, vacation, flight, day, eve) is built containing 6 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 author of this thesis randomly selected 100 appointment entries and determined whether each entry was a meeting. The system made the same decision as the author 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.

3.5 Chapter Summary

In this chapter, the two email collections (Enron and Avocado) and the Enron Phone Recording collection were introduced. To resolve the named mentions in

email and phone recording collection, the coverage of the general knowledge base Wikipedia for the mentions is studied. The results show that collection-specific person and organization knowledge bases are needed in order to resolve the mentions. Following the prior work of [30], a collection-specific person KB could be built by extracting the email addresses as potential entities. Then the method of constructing a collection-specific organization KB is introduced by extracting the domain names from email addresses as potential organization entities. The attributes, including the official website, Wikipedia page, and name variants of the entities are extracted through four sources (Google, Wikipedia, email Body and signature). As the linking targets for the meeting linking task, a collection-specific meeting KB is built by extracting the work-related appointment entries from the calendars. The general KB Wikipedia, the collection-specific person, organization and meeting KBs constitute the searching space for named mentions and meeting related email messages.

Chapter 4: Entity Linking for Email

In this chapter, an entity linking system is introduced to link three types of named mentions (person, organization, location) to the general knowledge base Wikipedia, and collection-specific knowledge bases. The efficacy of the feature groups are studied in detail for the task of person entity linking.[1]

Two email collections, Enron and Avocado, are used for the experimental study. The Enron email collection is used for the system design and development, and the Avocado email collection is used only for the efficacy testing. Following the TAC entity linking task, three types of named mentions are selected (person, organization, location). However, with more types of named mentions recognized (e.g., vehicle, facility, currency), the entity linking system has the potential to link them to the general knowledge base Wikipedia and other collection-specific knowledge bases if applicable. In this thesis, to resolve the person and organization mentions, separate corresponding collection-specific knowledge bases are used as the linking targets.

On the Enron and Avocado email collections, named mentions of three types are extracted from randomly selected email messages, manually linked to the knowl-

[1]This work has been published in Gao, N., Dredze, M., & Oard, D. W. (2017). Person entity linking in email with NIL detection. Journal of the Association for Information Science and Technology, 68(10), pages 2412-2424 [46].

edge bases, and used as ground-truth. The task of the entity linking system is automatically linking the named mentions to the referenced entities in the knowledge bases, described in Section 4.1. The entity linking system is a supervised machine learning system using a large set of features (Section 4.2). The most important part of the system is the feature design, which is introduced in Section 4.3. The evaluation metrics used for both entity linking and meeting linking tasks are introduced in Section 4.4. The test collection used in previous work and built for the task is introduced in Section 4.5. The experiment results are shown in Section 4.6 for all three types and in Section 4.7 for person linking.

4.1 Task Definition

This thesis focuses on linking three types of named mentions (person, organization, location) recognized from the email messages and transcripts of conversational telephone speech to three available knowledge bases (Wikipedia, collection-specific person KB, and collection-specific organization KB). For the entity linking task, manually recognized named mentions with identified types are provided as ground-truth named mentions. In this section, the formal definition of the task is given.

- **Knowledge base.** Let S be an email collection. Collection specific KBs $\{K_p, K_o\}$ are built for PER and ORG for corresponding collection-specific entities $\{E_p, E_o\}$ from the collection. Including the entities E_w in the knowledge base built from Wikipedia K_w, the entity search space is defined as $E := \{E_p \cup E_o \cup E_w\}$. Each entity from Wikipedia $e_w := f_w(N, T)$ is rep-

resented by the name variants N and the content of the Wikipedia page T. Each organization entity from the ORG KB $e_o := f_o(N, D, A)$ is represented by the name variants N that are extracted from different sources (e.g., Yahoo!), the domain address (e.g., yahoo) D, and the associated email addresses A in the corresponding email collection.

- **Query Named Mention.** Let $Q := \{q_i\}$ be a set of named mentions manually observed in a subset of email messages μ where $\mu \subset S$. We can extend the context of each named mention q_i by $q_i := f(P, E_i, T_i)$, where P is the type of the recognized named mention (PER, ORG or LOC), $E_i \subset E$ is a set of entities that participate in the email message μ_i, and T_i is a vector of words representing the content of message μ_i.

- **Collection-Specific Person Entity.** The collection-specific PER KB is built from the email collection S. Each person entity $e_p \in E_p$ is uniquely represented by $e_p := f_p(N, C, M')$, where N is the name variants of the entity, $C := \{c_{e_p,e_t}\}$ is the contact list of entity e_p in email collection, formed by a set of entities e_t that have been observed in the same email message with entity e_p. For each entity e_t in e_p's contact list, c_{e_p,e_t} is the frequency with which the two entities are observed in the same email message, and $M' \subset M$ is a set of email messages that contain entity e_p as a participant.

Given all the available sources for each entity $e \in E$ from the knowledge bases, and the extended context $q_i = f(P, E_i, T_i)$ of named mention q_i, the task is to identify the true referent entity e that named mention q_i is referring to, or return

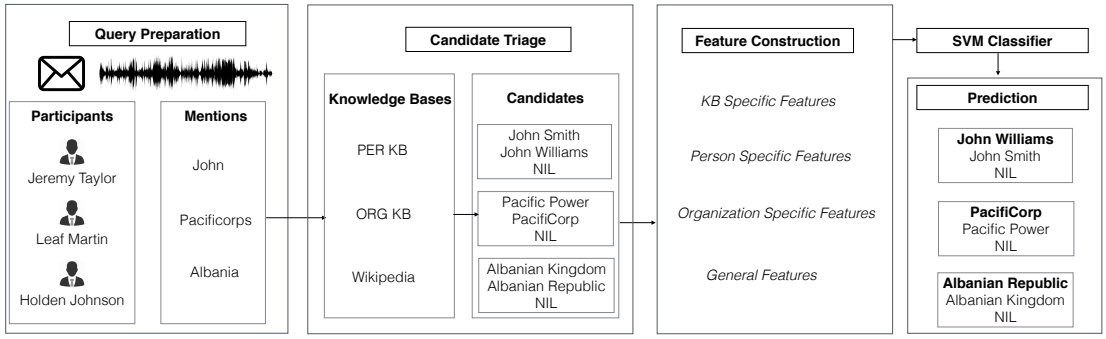

Figure 4.1: Framework of the multi-KB entity linking system.

ϕ (for NIL case) if the true referent is absent from all the KBs.

4.2 System Design

Figure 4.1 shows the framework of the proposed entity linking system. The feature-based supervised entity linking system is composed of four stages: query preparation, candidate triage, feature construction and prediction. In the query preparation stage, necessary contexts (e.g., metadata of the query email messages, content of the messages) are gathered for each named mention to be used in the linking stage. For each query named mention q_i, the context of the mention $q_i := f(P, E_i, T_i)$ is extracted. In the example in Figure 4.1, there are three named mentions in the query email message, the person mention Geir, the organization mention Pacificorps, and the location mention Albania. The participants in the conversation E_i are Jeremy Morris, Leaf Harasin and Holden Salisbury.

There are three KBs used as linking targets: Wikipedia and the collection-specific PER KB [30] and ORG KB built from emails introduced in Section 3.3. Each query mention could be referring to any of the entities in the knowledge base, or to

none of them (NIL). Triaging is a classification step targeting at narrowing down the number of candidates from all the entities in the knowledge bases to a relatively small set, and thus improving the ranking efficacy and efficiency of the next step. In the proposed system, there are multiple knowledge bases providing linking targets for different types of entities. Thus unlike prior work, the candidate triage step in this thesis first selects the proper KBs (e.g., Wikipedia and collection-specific organization KB for an organization named mention), and then selects candidate entities from those KBs.

In detail, the candidate triage step identifies possible candidates from the KBs for the query mentions based on a cascade of standard heuristics: (1) only the entities that match the mention type (PER, ORG or LOC) can be candidates. To be more specific, the candidates for PER mentions can only be extracted from the PER KB and Wikipedia. The candidates for ORG mentions candidates can only be extracted from the ORG KB and Wikipedia. The candidates for LOC mentions can only be extracted from Wikipedia; (2) exact string match; (3) match on initials (e.g., entity Imperial Irrigation District is a match for query mention IID); (4) fuzzy match in a way that the entity name contains all the words in the mention, or the mention contains all the words in the entity name (e.g., entity United States of America is a match for query mention United States); (5) fuzzy match in a way that the string of the entity name starts with the query mention (e.g., entity California is a match for query mention Cal); (6) ϕ is a candidate for all the queries to indicate that the true referenced entity may be absent from all available knowledge bases. The entities extracted following the aforementioned six steps are formed as the candidate set $\mathcal{E}_i$

of mention q_i.

The third step, feature construction generates a large set of features for each (*mention*, *candidate*) pair from the triage phase. These features are used to score the *candidate* for the given *mention* in the conversation. The features are organized into four groups for presentation purposes, including *General Features* considering the string match that have proven to be useful in prior work; *KB Specific Features* indicating the KB that the candidate is from, *Person-Specific Features* that utilize the contact behavior (from only person candidates), and the *Organization Specific Features* designed for only candidates from the ORG KB. The features are explicitly introduced in Section 4.3. All types of candidates share the same set of features.

Within those four groups of features, *Person-Specific Features* are explored to a degree in prior work for conversational content. In this thesis, new sources and new shapings of those features are developed and explored to improve person linking efficacy. *KB Specific Features* are novel features designed due to the specialized multiple knowledge base structure used in the system. *Organization Specific Features* are also novel features designed particularly to explore the best sources and features for the organization linking task, which have not been studied before.

For prediction, a Support Vector Machine (SVM) is used to rank the candidates based on the above features. The SVM regression model nu-SVR with a radial basis kernel from LibSVM [14] is applied. The top scoring candidate is the system's prediction, but the quality of the ranked list is evaluated as well. If treated as an entity linking task, the top prediction could be used for the purpose of understanding the content, thus Accuracy is chosen as an evaluation metric used in this thesis. If

treated as a ranking task, the predicted ranked list for each named mention could be used to support entity retrieval related tasks, and thus Mean Reciprocal Rank, which is introduced in detail below in Section 4.4, is also used as the evaluation metric.

4.3 Feature Design

In this section, the features $\mathbb{D} = \{\mathscr{D}_k\}$ used in the system are explicitly listed. Each feature $\mathscr{D}(q_i, e \in \mathscr{E}_i)$ is designed to reflect the probability that the candidate $e \in \mathscr{E}_i$ being the true referent for query mention q_i or the true referent is absent from the all the KBs, which referred to as the case of ϕ.

4.3.1 General features

General features are extracted for all types of query mentions, including features measuring if there is a string match or fuzzy match between the query and the name variants of the candidate entity:

$$\mathscr{D}(q_i, e) := |\{n \in N : n = q_i\}|. \tag{4.1}$$

A feature CANDIDATEISNIL is also built to indicate whether the candidate is the NIL candidate ϕ.

4.3.2 Person-specific features

There are a large set of features developed for the purpose of linking person named mentions to the collection-specific KB, organized into 12 groups according to the sources used.

GlobalNameVariantsGroup[Baseline] Features *nameMatch* and *fuzzyNameMatch* defined as

$$\mathscr{D}(q_i, e) := |\{n \in N : n = q_i\}| \tag{4.2}$$

are built to indicate if there is a match between the mention string q_i and the name variants N of the candidates. The username portion of all email addresses associated with an entity are also extracted as additional name variants. Two features *emailNameFieldMatch* and *fuzzyEmailNameFieldMatch* are created. Here *name* is a known name variant for the entity, and *emailNameField* is the username field of the email address (A in A@B). Since the collection-specific KB contains prior probabilities for each name variant, another set of features are built using those probabilities: *nameMatchProb*, *emailNameFieldMatchProb*, *fuzzyEmailNameFieldMatchProb*, and *fuzzyNameMatchProb*:

$$\mathscr{D}(q_i, e) := \sum_{n_i \in N} \left(\frac{\sum_{a \in A} f(a, n_i)}{\sum_{a \in A, n \in N} f(a, n)} \cdot |\{n_i : n_i = q_i\}|\right). \tag{4.3}$$

ParticipantsGroup. A set of 10 features extracts information from the header of the email message μ_j containing the query mention q_i and the corresponding email

thread $\hat{\mu}_j$. For each query mention q_i, features are set to check whether the header E_i contains the candidate e:

$$\mathscr{D}(q_i, e) := |e \cap E_i|. \tag{4.4}$$

By shaping the header set E_i using different fields, features *inHeader*, *inHeaderExceptFromField*, *inToField*, *inFromField*, *inCcField*, and *inBccField* are developed. Feature *inThreadHeader* uses the participants of the email thread by replacing the E_i in equation 4.4 with $\hat{E}_i$. Then real-valued versions of some features are created. The intuition is that if the email is sent to 100 people, a candidate matching one of the recipients would be modelled by normalized features as less informative. Therefore, features are created by normalizing the values by the number of participants: *inToFieldNormalized*, *inCcFieldNormalized*, *inBccFieldNormalized*.

$$\mathscr{D}(q_i, e) := \frac{1}{|E_i|} \cdot |e \cap E_i|. \tag{4.5}$$

MessageNameVariantsGroup[Baseline]. A set of 8 features is generated based on whether the name variants N of candidate entity E (or a substring from that name) appears in the email message μ_j containing query mention q_i: *candidateInEmail*, *fuzzyCandidateInEmail*

$$\mathscr{D}(q_i, e) := |\{n \in N : n \cap T_i \neq \emptyset\}|. \tag{4.6}$$

Features are created based on whether mention q_i appeared in the email thread $\hat{\mu}_j$: *candidateInThread*, *fuzzyCandidateInThread* by replacing the T_i in the formula with $\hat{T}_i$. In collection-specific person KB, an entity can have several known name variants for which prior probabilities can be computed (e.g., Entity 20216 is mentioned by the name Jim 38% of the time). Therefore another set of real valued versions of these features are created using the prior probability of the candidate being mentioned by the name: *candidateInEmailProb*, *fuzzyCandidateInEmailProb*

$$\mathscr{D}(q_i, e) := \sum_{n_i \in N} \left(\frac{\sum_{a \in A} f(a, n_i)}{\sum_{a \in A, n \in N} f(a, n)} \cdot |\{n_i : n_i \cap T_i \neq \emptyset\}|\right). \tag{4.7}$$

Similarly, two features are created using the email thread *candidateInThreadProb* and *fuzzyCandidateInThreadProb* by replacing the T_i in the formula with $\hat{T}_i$.

ContactFrequencyGroup. 6 features are used to capture the aspects of the larger social context associated with the candidate. Let $e_i \in E_i$ be a participant of the email message, $e \in \mathscr{E}_i$ be a candidate entity, $C_{e_i,e}$ be the contact frequency between e_i and e, features are built by summing the contact frequency between e and the entities in the header E_i (*inContacts*):

$$\mathscr{D}(q_i, e) := \sum_{e_i \in E_i} C_{e_i,e}, \tag{4.8}$$

inFromContacts only uses the contact frequency between the sender and the candidate e, *inContactsThread* sums the contact frequency between the candidate entity e_i and the participants in the email thread $\hat{E}_i$ by replacing the E_i in the equation.

The sending behavior of the sender is also considered by measuring the probability that the sender emailed the candidate given that the sender emailed any person with the same name variant *contactProbability*:

$$\mathscr{D}(q_i, e) := \sum_{e_i \in E_i} \frac{C_{e_i,e}}{\sum_{e_j \in E}\{C_{e_i,e_j} : N_j \cap q_i \neq \emptyset\}}. \tag{4.9}$$

Even if two entities appeared in the same email message, the confidence that the two entities are acquaintances decreases when the number of participants gets larger. So features *inContactsNormalized* and *inContactsThreadNormalized* are developed to adjust the contact list with probabilities. The contact frequency of $C_{e_i,e}$ is defined as $\sum_{M'_k \in \{M'_{e_i} \cap M'_e\}} \cdot \frac{2}{|E_k|}$ where E_k is the participants of $M'_k \in \{M'_{e_i} \cap M'_e\}$. So the feature *inContactsNormalized* is defined as

$$\mathscr{D}(q_i, e) := \sum_{e_i \in E_i} \sum_{M'_k \in \{M^i_{e_i} \cap M'_e\}} \frac{2}{|E_k|}. \tag{4.10}$$

The feature *inContactsThreadNormalized* replaces the E_i in the formula with $\hat{E}_i$.

CommunicationCohortGroup. The entities E_i in a message header (or in the headers of messages in a thread $\hat{E}_i$) can be seen to form a social group. To measure how related the candidate e is to this group, all email messages $M' := \{M'_j \cap M'_e\}$ between the sender e_j and the candidate e are used as evidence. For each email message $m \in M'$, let E_m be its participants. Features are built by measuring the social context similarity between the sender and the candidate based on the Jaccard

similarity coefficients of the participant group of the message E_i with every other group E_m:

$$\mathscr{D}'(q_i, e) := \frac{|E_i \cap E_m|}{|E_i \cup E_m|} \cdot |\{m_k \in M' : (E_m = E_k) \cap (E_m \neq E_i)\}|. \tag{4.11}$$

Features are built for the maximum *socialContextMaxFrequency*

$$\mathscr{D}(q_i, e) = \mathrm{MAX}_m D'(q_i, e) \tag{4.12}$$

to detect the most similar communication group other than E_i that the sender and candidates are participating in, the mean *socialContextMeanFrequency*

$$\mathscr{D}(q_i, e) := \frac{\sum_m D'(q_i, e)}{|M'|} \tag{4.13}$$

to detect the average similarity between E_i and the communication groups including the sender and candidate, the total *socialContextSumFrequency* $\mathscr{D}(q_i, e) = \sum_m \mathscr{D}'(q_i, e)$, and a variant of the maximum *socialContextMax* by binarizing the frequency of observing same communication group

$$\mathscr{D}(q_i, e) := \mathrm{MAX}_m \frac{|E_i \cap E_m|}{|E_i \cup E_m|}. \tag{4.14}$$

TopicalContextGroup[Baseline]. 24 features are used to measure the content similarity between the body of the email message T_i in which the mention q_i was found, and the corresponding fields of all email messages M' that contain the can-

didate e as a sender or recipient. For the mention context, either the words in the *Email* T_i, or all the words in the *Thread* $\hat{T}_i$ are used as keywords. For the candidate context, all the email messages M' the person sent or received are used as the search space. Using the words $\{t\}$ in the email T_i or thread $\hat{T}_i$ as keywords, let $\mathcal{B}(t, M')$ be the term frequency of t in email messages M', the feature is defined as

$$\mathcal{D}(q_i, e) := TF * IDF * Norm, \tag{4.15}$$

where there are three ways to calculate the TF by using:

$$TF := \begin{cases} \mathcal{B}(t, M') \\ 1 + \log(1 + \mathcal{B}(t, M')) \\ 1 + \log(1 + \log(1 + \mathcal{B}(t, M'))) \end{cases} . \tag{4.16}$$

The IDF of the keywords t is calculated based on the whole collection *CollIDF*, or the *PerIDF* calculated based on the email messages that contain the candidate:

$$IDF := \begin{cases} \log \frac{1}{|\{m \in M : t \in T_m \neq \emptyset\}|} \cdot |M| \\ \log \frac{1}{|\{m \in M' : t \in T_m \neq \emptyset\}|} \cdot |M'| \end{cases} , \tag{4.17}$$

where T_m is a vector of words representing the body text of email message m. Since the email collections are incomplete for most of the entities in the KB, a low similarity score could be caused by the data sparsity rather than by the dissimilarity of their topical interests. Therefore, another set of features is built to *Norm* the

similarity score by the number of email messages containing the candidate $|M'|$:

$$Norm := \begin{cases} 1 \\ \\ \frac{1}{|M'|} \end{cases} . \tag{4.18}$$

By using two mention contexts T_i or $\hat{T}_i$, three TF calculations, two IDF methods, and normalization or not, there are 24 features in total for this group.

OrganizationGroup. Features indicating if the entity has an email address suggesting employment by the company whose emails are collected (Enron, Avocado): *isEmployee* is built. The sender being a company employee could indicate that the mention refers to company employee (and thus likely being in the KB) so we set *insideSender* in such cases; a similar feature is set when the sender and all recipients are all employees of the company: *allEmployee*. When introducing someone from another organization, the sender may write something like "I talked with John from Reliant Energy yesterday". A feature *potentialOtherORG* is set to recognize sentence structures of "*mention* from/work for/at *caps-initial*".[2]

NilDetectionGroup. The final four features are intended principally to improve performance on NIL detection. In the experiments, the effect of adding feature groups is analyzed, but because the focus on NIL detection is novel, each feature in the *NilDetectionGroup* is analyzed separately.

- *privateContext.* In the initial inspection of unresolvable referents in Elsayed's

[2] *caps-initial* indicates any term that begins with a capital letter.

Enron collection, it is noticed by the author that some mentions of people who do not appear in the KB (i.e., NILs) were family members. Therefore the words in the "family" category (e.g., father, dad) of the Linguistic Inquiry and Word Count (LIWC) dictionary [128] are used to detect private context. The feature *privateContext* is set when a word on that list is present within three lines above or below the mention.

- *entertainContext.* The unresolvable referents in Elsayed's Enron collection also suggests that some NILs in that collection are references to celebrities. A second feature is constructed, *entertainContent*, if a word matching the LIWC dictionary leisure category (which includes, for example, football and baseball) appears within three lines above or below the mention.

- *unknownSender.* Messages that are received from unknown senders are perhaps less likely to mention known individuals. This intuition is reflected by the construction of a binary feature *unknownSender*: if no candidate (for the particular query mention) has ever (in the email collection we are processing) sent an email message to or received a message from the sender, the *unknownSender* feature is set to be true for every candidate. Otherwise the *unknownSender* feature is set to be false for every candidate.

- *unknownFullName.* When introducing someone new in a conversation, it is a common practice to mention the full name of that person. The system then therefore looks in the *Path* from the query email message *To* the *Root* of the thread (PTR) for a multi-token name mention that matches the query mention

as a known variant (e.g., a mention of "James Foster" in the PTR would match a mention "James" in email). When such a match exists, the system checks the KB to see if there is at least one matching entity with that multi-token name. If not, the feature *unknownFullName* is set to true to indicate that the reference matches an unknown full name earlier in the PTR.

4.3.3 Organization-specific features

A group of features are designed for only organization mentions. The collection-specific organization KB is built by extracting all the domain names from the email addresses in the email collection as candidate organization entities. For each entity in the KB $e_o := f_o(N, D, A)$, there are name variants N of the organization extracted from Wikipedia, Google Search, email message body and signature. For each entity, there are also a set of email addresses A in the email collection that use the particular domain, potentially indicating the organization of the email address owner. To fully use all the information in the collection-specific KB, the features are built including the number of email addresses that use the current candidate organization domain

$$\mathscr{D}(q_i, e) := |A|; \tag{4.19}$$

the number of total levels D of the entity domain name (e.g., the domain level for store.yahoo.com is 3); and the level that there is a string match between the organization domain and query mention (e.g., there is a string match between query mention Yahoo and organization domain store.yahoo.com at level 2).

4.3.4 KB specific features

KB specific features include features indicating if the current candidate entity is from a collection-specific KB or the general KB built from Wikipedia (*collection-SpecificKB*)

$$\mathscr{D}(q_i, e) := \begin{cases} 0, \ if \ e \in \{E_p, E_o\} \\ 1, \ if \ e \in E_w \end{cases} . \tag{4.20}$$

For candidate entities from Wikipedia, the number of links point to the candidate entity is included as a feature *inLinks*. A feature (*wikiTitle*) is built to indicate if the query mention has an exact match of a Wikipedia page title, however, the entity described by the Wikipedia page is not included in the KB.

4.4 Evaluation Metrics

For each query mention q_i, the set of candidate entities $\mathscr{E}_i$ will be sorted by the possibility they are the true referent according to the system's judgement. If the true referent is in the candidate set $\mathscr{E}_i$, let r_i be its rank in the sorted list. If the true referent is not in $\mathscr{E}_i$, $r_i = +\infty$. Two metrics are used in this thesis to evaluate entity linking performance: the accuracy over all query mentions in Q:

$$\frac{1}{|Q|} \cdot |\{q_i \in Q : r_i = 1\}|, \tag{4.21}$$

and mean reciprocal rank (MRR):

$$\frac{1}{|Q|} \cdot \sum_{q_i \in Q} \frac{1}{r_i}. \tag{4.22}$$

For all the experiments in this thesis, both accuracy and MRR are used to report the efficacy of the systems, while only MRR is used in feature analysis.

4.5 Test Collections

For the Enron collection, two sets of ground truth entity linking annotations are available. Namata produced a set of 78 ground truth annotations for Non-NIL mentions that refer to Enron employees [31]. Elsayed later produced a second test collection by randomly sampling mentions [31]. This yielded a set of 467 Non-NIL mentions. Elsayed's annotators were unable to manually link an additional 112 of the randomly sampled mentions to any person in the KB, which can be divided into two categories: (1) a reference to a person who would not reasonably be expected to have sent or received an email message that is in the collection (i.e., a NIL mention), or (2) a reference that the annotator was unable to resolve due to insufficient time or insufficient understanding of the implicit context, which we refer to as "unresolvable". Therefore, two independent annotators were asked to mark each of Elsayed's 112 unresolved mentions as NIL if there was good reason to believe that it was actually a NIL reference. A total of 45 of the 112 unresolved mentions were marked as NIL by at least one of the two annotators. The agreement on this two-way classification task (NIL or unresolvable), measured by Cohen's Kappa, was

0.575. These 45 NIL annotations were added to Elsayed's 467 Non-NIL annotations to produce a unified test set (which referred to as Elsayeds evaluation set) that contains a total of 512 annotations.

Table 4.1 shows a new evaluation set built for the task of entity linking for three types of mentions to the multiple knowledge bases for Enron emails. Named mentions are extracted automatically from 113 randomly selected email messages by using the Illinois Named Entity Tagger, and linked to the general knowledge base Wikipedia and collection-specific knowledge bases by six independent annotators.

To create the test collection for person entity linking on the Avocado email collection, 250 single-token named mentions are selected from randomly selected email messages, and one independent annotator was asked to make annotations, using a simple search system to find messages that could provide useful context based on content or person name. The simple annotation platform additionally provided the annotator with a KB browser, through which the annotator could determine the most frequent contacts of each entity. This resulted in 148 mentions annotated as Non-NIL, 56 mentions annotated as NIL, and 46 mentions annotated as unresolvable.

To characterize inter-annotator agreement, the author of this thesis independently annotated 20 randomly selected mentions from the same set of 250. In that set of 20, the independent annotator had marked two mentions as unresolvable; the author agreed in one instance, and made a Non-NIL resolution in the other. On the remaining 18 mentions, 4 were marked as NIL by the independent annotator, the author agreed on three of those four, and made a Non-NIL annotation on the fourth.

The remaining 14 annotations received Non-NIL annotations from both annotators, and those annotations were identical in 12 of the 14 cases. Thus the overall agreement on the 14 annotations judged by the independent annotator to be Non-NIL was $12/14 = 86\%$. This level of agreement is consistent with the 80% agreement on Non-NIL mentions reported by Elsayed for the Enron collection [29].

Then on the same set of randomly selected email messages, the organization and location mentions were automatically extracted by using the Illinois Named Entity Tagger and linked to the collection-specific organization knowledge base and Wikipedia manually by the author of this thesis. Table 4.2 shows the annotation results, used as the ground-truth entity linking annotations for the Avocado email collection. Comparing with the Enron email collection, there are fewer links from person named mentions to Wikipedia on the Avocado email collection. One cause for this difference is that there are more Enron employees frequently mentioned (e.g., the CEO Kenneth Lay, Jeffery Skilling) on Wikipedia due to the Enron scandals.

For the entity linking task for all three types, there are three KBs used as the linking targets: (1) the TAC 2008 KBP Reference Knowledge Base (which contains PER, ORG and LOC entities); (2) a collection-specific PER KB [30] containing 124,475 person entities; and (3) and a collection-specific ORG KB containing 23,008 organization entities. Both the collection-specific person, organization knowledge bases are built from the Enron email collection.

Table 4.1: Human annotations for the linking on Enron Emails.

	All	PER KB	ORG KB	Wikipedia	NIL
PER	150	53	0	49	52
ORG	134	0	75	63	39
LOC	181	0	0	116	65
Total	465	53	75	228	156

Table 4.2: Human annotations for the linking on Avocado Emails.

	All	PER KB	ORG KB	Wikipedia	NIL
PER	202	148	0	13	43
ORG	72	0	52	36	13
LOC	85	0	0	56	29
Total	359	148	52	105	85

4.6 Entity Linking Results for All Types

Table 4.3 shows the entity linking results for all three types of mentions for the Enron email collection. Table 4.4 shows the results for the Avocado email collection. Since there is no prior work on the same task, *Random* ranks all the candidates for each named mention randomly, this is used as the lowest baseline. *Baseline* adds all the *Baseline* feature groups to the entity linking system, used as a higher baseline. By using the Illinois Named Entity Tagger, both the named mentions and their predicted types (person, organization, location) are detected automatically. *All features (Separate models)* adds in all the features to the system, and trains separate SVM models for each mention type. *All features (One model)* also uses all the features, but trains a single SVM model for all types of mentions.

As can be seen, training separate models performs slightly better than training one single model for all types in most of the cases. It is generally harder to predict the NIL mentions than resolving Non-NIL mentions, especially for location mentions.

Table 4.3: Entity linking for all mentions, Enron email collection.

		Non-NIL	NIL	All
PER	*Random*	0.102	0.153	0.12
	Baseline	0.466	0.346	0.692
	All features (Separate models)	**0.744**	**0.756**	**0.748**
	All features (One model)	0.699	0.701	0.700
ORG	*Random*	0.116	0.391	0.387
	Baseline	0.333	0.209	0.565
	All features (Separate models)	**0.758**	**0.913**	**0.812**
	All features (One model)	0.741	0.850	0.781
LOC	*Random*	0.224	0.545	0.340
	Baseline	0.230	0.189	0.303
	All features (Separate models)	**0.882**	0.593	0.778
	All features (One model)	0.874	**0.668**	**0.800**
All	*Random*	0.212	0.340	0.12
	Baseline	0.336	0.246	0.514
	All features (Separate models)	**0.819**	0.727	**0.778**
	All features (One model)	0.779	**0.731**	0.762

One possible reason is that the Wikipedia knowledge base used in the experiments is derived from the infoboxes of Wikipedia pages. Some of the referenced locations are absent from the knowledge base due to the lack of infoboxes while LDC[3] generating the dataset. For a location named mention "Santa Clara", it is difficult for the system to predict that the true reference "Santa Clara, California" is absent from the knowledge base (the NIL case) while there are other candidates available (e.g., Santa Clara, Texas). A new version of the Wikipedia knowledge base, or using the Wikipedia pages directly as the linking target, might be the solution to this problem.

4.7 Entity Linking Results for Person

In this section, the evaluation results are explicitly analyzed for the task of linking person named mentions to a collection-specific person knowledge base.

[3] `https://catalog.ldc.upenn.edu/LDC2014T16`

Table 4.4: Entity linking for all mentions, Avocado email collection.

		Non-NIL	NIL	All
PER	*Random*	0.087	0.185	0.112
	Baseline	0.284	0.490	0.336
	All features (Separate models)	**0.767**	0.853	**0.789**
	All features (One model)	0.727	**0.860**	0.761
ORG	*Random*	0.185	0.0	0.151
	Baseline	0.550	0.319	0.509
	All features (Separate models)	**0.829**	**0.667**	**0.799**
	All features (One model)	0.815	**0.667**	0.788
LOC	*Random*	0.214	0.307	0.243
	Baseline	0.447	0.534	0.475
	All features (Separate models)	**0.911**	0.567	0.802
	All features (One model)	**0.911**	**0.590**	**0.809**
All	*Random*	0.133	0.195	0.149
	Baseline	0.371	0.481	0.399
	All features (Separate models)	**0.815**	0.727	**0.794**
	All features (One model)	0.783	**0.759**	0.776

4.7.1 Non-NIL Results

Research results have been reported on the Elsayed's collection from two prior systems **Elsayed** and **Xu**. In both cases, results were reported only for Elsayed Non-NIL mentions, and Elsayed's system used the Namata collection as training. So Table 4.5 shows the comparison of the proposed system entity linking results with those two prior results using the same setting: training on Namata Non-NIL; testing on Elsayed Non-NIL. As Table 4.5 shows, the results of Xu's unsupervised system is not comparable to the results of the two supervised techniques: Elsayed's and the proposed system. Also the entity linking system achieves a 20% error reduction in Mean Reciprocal Rank (MRR) and a 12% error reduction in accuracy over Elsayed's system, both of which are statistically significant by a paired, two-tailed t-test ($p < 0.05$).

Table 4.5: Train: Namata Non-NIL; test: Elsayed Non-NIL.

	Xu	Elsayed	Gao (Our system)
MRR	0.667	0.785	**0.827**
Accuracy	0.564	0.739	**0.771**

Table 4.6: Entity linking system, train/test: Elsayed.

Train	Non-NIL	Both Non-NIL and NIL		
Test	Non-NIL	Non-NIL	NIL	Both
MRR	0.822	0.817	0.752	0.811
Accuracy	0.777	0.740	0.713	0.738

Table 4.7: Entity linking system, train: Elsayed, test: Avocado.

Train	Non-NIL	Both Non-NIL and NIL		
Test	Non-NIL	Non-NIL	NIL	Both
MRR	0.884	0.898	0.360	0.756
Accuracy	0.871	0.860	0.130	0.667

To further explore the performance of the proposed system, Table 4.6 shows results by 2-fold cross-validation on a random split of the Elsayed collection. Table 4.7 shows results by training on Elsayed collection and testing on our Avocado collection. Table 4.8 shows the 2-fold cross-validation result by training and testing on Avocado collection. Notice that the Avocado collection is used only for performance validation. None of the system features are adjusted to the new collection. Cross-validation can sometimes yield artificially good results because of unusual similarities between the training and test data, but comparing Table 4.5 to the first result column of Table 4.6 shows that such an effect is not observed in this case (compare MRR of 0.827 to 0.822, and accuracy of 0.771 to 0.777). As a conclusion, reporting cross-validation results for the Elsayed collection is reasonable.

Table 4.8: Entity linking system, train/test: Avocado.

Train	Non-NIL	Both Non-NIL and NIL		
Test	Non-NIL	Non-NIL	NIL	Both
MRR	0.934	0.926	0.879	0.912
Accuracy	0.915	0.901	0.806	0.873

4.7.2 NIL Results

Table 4.7 and 4.8 show the results of the entity linking system for training with Non-NIL and NIL mentions together. Since the entity linking system is the first that is able to detect NIL mentions, there are no state-of-the-art baselines to compare with. Comparing the first and second columns in the tables indicates that training on both Non-NIL and NIL mentions yields results similar to training only on Non-NIL mentions, at least as measured by NIL MRR (compare 0.822 to 0.817 in Table 4.6, 0.884 to 0.898 in Table 4.7, and 0.934 to 0.926 in Table 4.8), which demonstrates that the NIL detection doesn't adversely affect the results of our system on Non-NIL mentions.

NIL detection yields high performance when training and testing on the same collection: NIL MRR 0.752 for train/test on Elsayed, and 0.879 for train/test on Avocado. However, when training on Enron and testing on Avocado, there is a significant drop in NIL MRR results compared with train/test on Avocado (from 0.879 to 0.360). Further analysis shows that this is because the characteristics of NIL mentions are different for different collections, yielding different learned models for NIL detection features. On Elsayed, a substantial number (39%) of the mentions judged as NIL were made when assessors encountered references to what seemed to

them to be family members or friends. For example, a message includes *We have two younger boys...Alex and Eddie* resulted in *Alex* and *Eddie* being judged as NIL. However, on Avocado, most of the NIL mentions are judged as NIL because of the cues in the email indicated to the annotator that the referent entity is absent from our current email collection. For example, the mention *Matthew* in the context *Matthew is moving to Houston* was judged as NIL because none of the candidate entities in our KB seemed to the annotator to be known to the sender or to any recipients.

4.7.3 Feature Group Significance Tests

Table 4.9 lists feature groups designed for linking person named mentions to collection-specific person KB and the corresponding features in each group. The feature *candidateIsNIL* is included in the analysis. The related work is listed if the same or similar design is observed for each feature. The novel features are also divided into two groups: novel evidence or novel shaping of features. Type shows if the current feature is Boolean or numeric.

Most of the features in four groups (*MessageNameVariantsGroup*, *GlobalNameVariantsGroup*, *TopicalContextGroup*, *candidateIsNil*) are similar in shape or source to the features used in prior work, which are combined and used as the **Baseline**. In Table 4.10 and 4.11, the efficacy of the four feature groups in Baseline are measured by MRR on Non-NIL, NIL and Overall named mentions on different collections. Then for the other feature groups that are novel in design or source

(e.g., *ParticipantsGroup*), the MRR is reported when using both **Baseline** and the particular feature group (i.e., Baseline + *ParticipantsGroup*). The MRR values are emphasized in bold if the improvement by adding in a particular feature group is statistically significant ($P < 0.05$) when evaluated by two-tailed paired t-test. For example, when training and testing on Enron, by adding the *ParticipantsGroup* to the Baseline, the Non-NIL MRR improves significantly from 0.409 to 0.642, which confirms its contribution to linking the Non-NIL named mentions.

The three feature groups originally designed for Non-NIL named mentions (i.e., *ParticipantsGroup*, *ContactFrequencyGroup*, and *CommunicationCohortGroup*) are shown to be useful for Non-NIL mentions in all four collection settings. They further improve the efficacy for NIL mentions on the Avocado collection, but not on the Enron collection. The feature group *OrganizationGroup* was initially designed for NIL mentions by recognizing the referents outside Enron. However, the experiments show that *OrganizationGroup* features are more effective in recognizing the referents that are inside Enron (Non-NIL), especially on Elsayed collection. *privateContext* is the strongest NIL detection feature when train/test on Elsayed. *entertainContext* and *unknownFullName* are the strongest features when training and testing on Avocado. When training and testing on different collections, the aforementioned *NilDetectionGroup* performs relatively poorly due to the different properties of different collections.

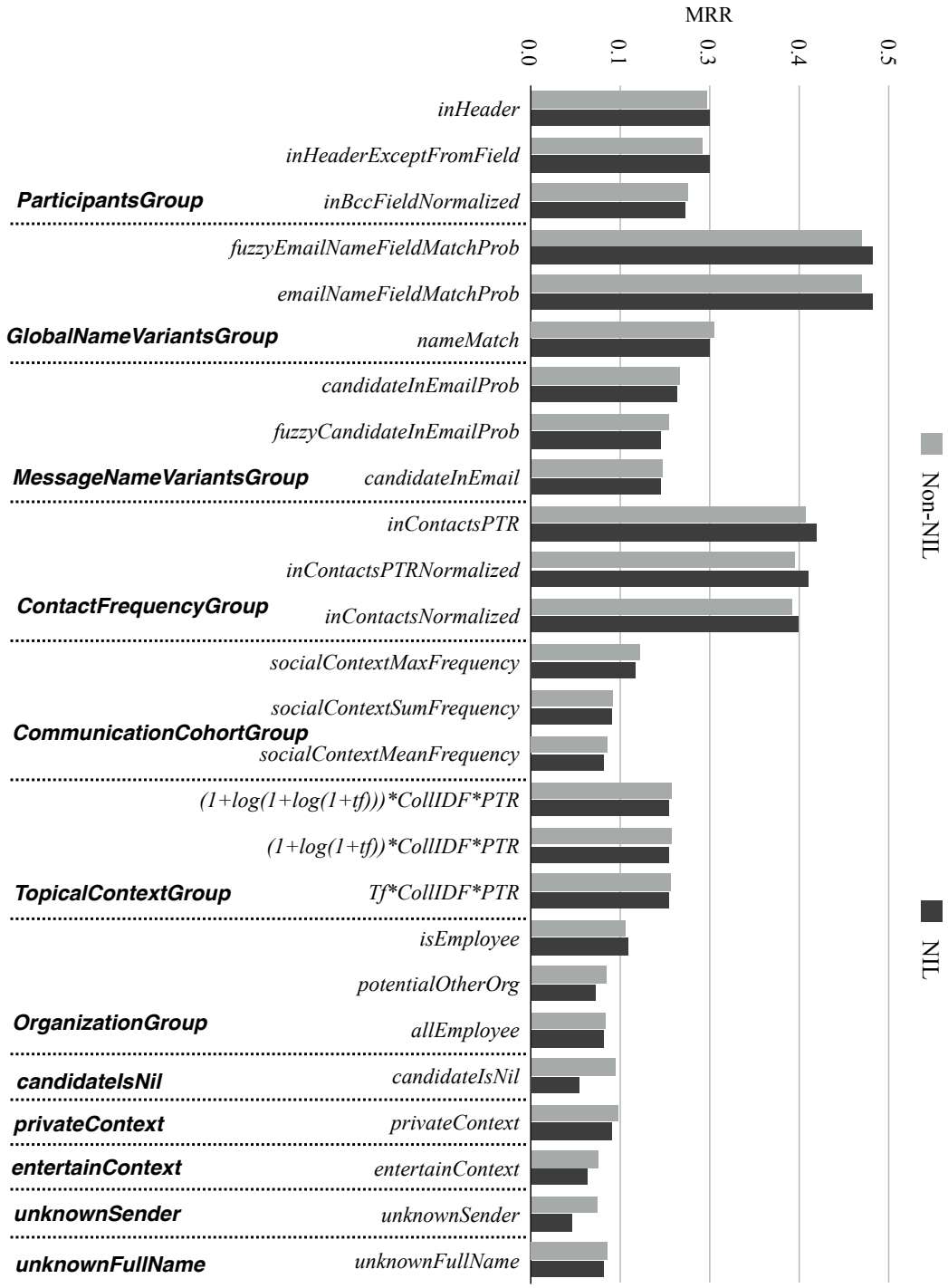

Figure 4.2: Single feature analysis on Elsayed collection.

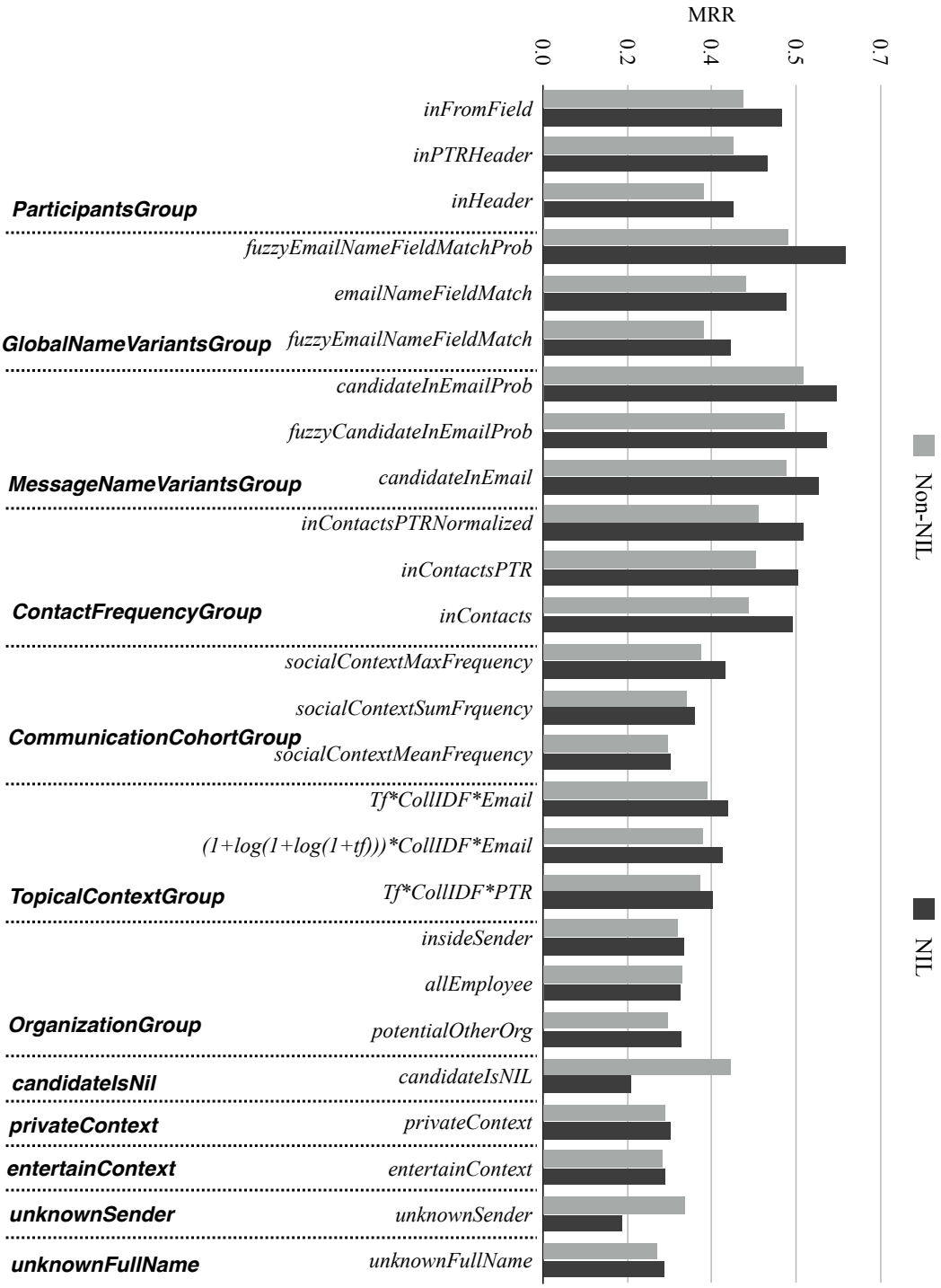

Figure 4.3: Single feature analysis on Avocado collection.

4.7.4 Single Feature Analysis

Figure 4.2 and 4.3 show the single feature efficacy by training/testing on the same email collection (Elsayed in Figure 4.2, Avocado in Figure 4.3). For evaluation, only one feature is added to the Feature Construction stage at a time. The figures show the efficacy for the top 3 features (with highest overall MRR) in each feature group and also, in particular, all the features in *NilDetectionGroup* since NIL detection is one of the focuses of this thesis. The light grey bars represent the Non-NIL MRR for each features, and the dark grey bars represent the NIL MRR.

For both Non-NIL and NIL, the best single features come from the *GlobalNameVariantsGroup* and *ContactFrequencyGroup* for Elsayed. All 3 of the best features in the *ContactFrequencyGroup* are novel shaping features proposed in this thesis, two of which (*inContactsPTR* and *inContactsPTRNormalized*) are also the best performing features in the *ContactFrequencyGroup* on the Avocado collection. This observation is consistent with the results in Table 4.10 that the overall MRR improves the most by adding in the *ContactFrequencyGroup* features into the **Baseline** on Elsayed collection. Out of the 24 features in the *TopicalContextGroup*, the only feature that has been used in previous work is *TF*CollIDF*Email*, which is reported as the best single feature in the group on the Avocado collection. There is no significant difference between the three proposed ways of calculating TF in our experiments. However, all the best features in the *TopicalContextGroup* calculate the IDF based on the whole collection rather than on entity-specific content. One explanation could be that most of the candidates only participate in a few email

messages, even less if there is any as a writer. As a result, the *PerIDF* calculated based on those participated messages might not accurately represent the language style of certain entities.

Within all the *NilDetectionGroup* features, *privateContext* shows the best NIL MRR as a single feature on Enron, which also contributes most to the NIL detection when added to the **Baseline**, as shown in Table 4.10. On Avocado, the best features in the NIL Detection group are *privateContext* and *UnknowFullName*. The NIL detection features are not the best performing singe features for either Non-NIL or NIL, however, they improve the NIL MRR significantly when added to the **Baseline**.

4.8 Chapter Summary

This chapter introduced an entity linking system for linking mentions of named mentions found in the body text of email messages to knowledge bases. The entity linking system is the first to consider the task of NIL mention detection in email. NIL detection accuracy results that are comparable to the best results previously reported for Non-NIL mentions is achieved by designing new features that are specifically motivated by the NIL detection task. The results show that inclusion of these features does not have any material negative effect on linking accuracy for Non-NIL mentions. NIL mentions comprise a substantial portion on both test collections, so NIL detection will be an important part of many practical applications of this technology. The entity linking system also achieved a 20% error reduction over the best previously reported results for linking Non-NIL mentions by combining both

existing and novel features using what is now a rather standard machine learning framework that had not yet been used for entity linking at the time the earlier work was performed.

Table 4.9: Feature Novelty.

Feature Group	Type	Feature	Related Work	Novelty
ParticipantsGroup	Boolean	*inHeader*	Xu [137]	
	Boolean	*inFromField*		shaping
	Boolean	*inHeaderExceptFromField*		Shaping
	Boolean	*inToField*		Shaping
	Boolean	*inCcField*		Shaping
	Boolean	*inBccField*		Shaping
	Numeric	*inToFieldNormalized*		Shaping
	Numeric	*inCcFieldNormalized*		Shaping
	Numeric	*inBccFieldNormalized*		Shaping
	Numeric	*inPTRHeader*	Elsayed [31]	
MessageNameVariantsGroup	Numeric	*candidateInEmail*	Elsayed	
	Numeric	*fuzzyCandidateInEmail*	Elsayed	
	Numeric	*candidateInPTR*	Elsayed	
	Numeric	*fuzzyCandidateInPTR*	Elsayed	
	Numeric	*candidateInEmailProb*	Elsayed	
	Numeric	*fuzzyCandidateInEmailProb*	Elsayed	
	Numeric	*candidateInPTRProb*	Elsayed	
	Numeric	*fuzzyCandidateInPTRProb*	Elsayed	
GlobalNameVariantsGroup	Numeric	*nameMatch*	McNamee [89]	
	Numeric	*fuzzyNameMatch*	McNamee	
	Numeric	*emailNameFieldMatch*	Elsayed	
	Numeric	*fuzzyEmailNameFieldMatch*	Elsayed	
	Numeric	*nameMatchProb*	Elsayed	
	Numeric	*emailNameFieldMatchProb*	Elsayed	
	Numeric	*fuzzyEmailNameFieldMatchProb*	Elsayed	
	Numeric	*fuzzyNameMatchProb*	Elsayed	
ContactFrequencyGroup	Boolean	*inContacts*	Diehl [24]	
	Numeric	*inContactsNormalized*		Shaping
	Boolean	*inFromContacts*	Diehl	
	Boolean	*inContactsPTR*		Shaping
	Numeric	*inContactsPTRNormalized*		Shaping
	Numeric	*contactProbability*	Elsayed	
CommunicationCohortGroup	Numeric	*socialContextMaxFrequency*		Shaping
	Numeric	*socialContextMeanFrequency*		Shaping
	Numeric	*socialContextSumFrequency*		Shaping
	Numeric	*socialContextMax*		Shaping
TopicalContextGroup	Numeric	*Tf*CollIDF*Email*	McNamee	
	Numeric	*Tf*CollIDF*Email*Norm*		Shaping
	Numeric	*Tf*CollIDF*PTR*		Shaping
	Numeric	*Tf*CollIDF*PTR*Norm*		Shaping
	Numeric	*Tf*PerIDF*Email*		Shaping
	Numeric	*Tf*PerIDF*Email*Norm*		Shaping
	Numeric	*Tf*PerIDF*PTR*		Shaping
	Numeric	*Tf*PerIDF*PTR*Norm*		Shaping
	Numeric	*(1+log(1+tf))*CollIDF*Email*		Shaping
	Numeric	*(1+log(1+tf))*CollIDF*Email*Norm*		Shaping
	Numeric	*(1+log(1+tf))*CollIDF*PTR*		Shaping
	Numeric	*(1+log(1+tf))*CollIDF*PTR*Norm*		Shaping
	Numeric	*(1+log(1+tf))*PerIDF*Email*		Shaping
	Numeric	*(1+log(1+tf))*PersIDF*Email*Norm*		Shaping
	Numeric	*(1+log(1+tf))*PerIDF*PTR*		Shaping
	Numeric	*(1+log(1+tf))*PerIDF*PTR*Norm*		Shaping
	Numeric	*(1+log(1+log(1+tf)))*CollIDF*Email*		Shaping
	Numeric	*(1+log(1+log(1+tf)))*CollIDF*Email*Norm*		Shaping
	Numeric	*(1+log(1+log(1+tf)))*CollIDF*PTR*		Shaping
	Numeric	*(1+log(1+log(1+tf)))*CollIDF*PTR*Norm*		Shaping
	Numeric	*(1+log(1+log(1+tf)))*PerIDF*Email*		Shaping
	Numeric	*(1+log(1+log(1+tf)))*PerIDF*Email*Norm*		Shaping
	Numeric	*(1+log(1+log(1+tf)))*PerIDF*PTR*		Shaping
	Numeric	*(1+log(1+log(1+tf)))*PerIDF*PTR*Norm*		Shaping
OrganizationGroup	Boolean	*insideSender*		Evidence
	Boolean	*allEmployee*		Evidence
	Boolean	*isEmployee*		Evidence
	Boolean	*potentialOtherOrg*		Evidence
NilDetectionGroup	Boolean	*privateContext*		Evidence
	Boolean	*entertainContext*		Evidence
	Boolean	*candidateIsNil*	McNamee	
	Boolean	*unknownSender*		Shaping
	Boolean	*unknowFullName*		Shaping

Table 4.10: Efficacy of adding feature groups to the baseline features, measured by MRR, test on Elsayed.

Feature Groups	Train/Test on Elsayed			Train on Avocado, Test on Elsayed		
	Non-NIL	NIL	Overall	Non-NIL	NIL	Overall
Baseline	0.409	0.488	0.416	0.415	0.595	0.431
ParticipantsGroup	**0.642**	0.233	**0.606**	**0.627**	0.107	**0.581**
ContactFrequencyGroup	**0.750**	0.130	**0.696**	**0.736**	0.112	**0.681**
CommunicationCohortGroup	**0.542**	0.084	**0.502**	**0.515**	0.083	**0.477**
OrganizationGroup	**0.498**	0.209	**0.473**	**0.490**	0.108	0.456
privateContext	0.440	**0.611**	**0.455**	**0.457**	0.195	0.434
entertainContext	0.450	0.266	0.434	**0.465**	0.274	0.448
unknownSender	0.467	0.414	0.462	**0.463**	0.141	0.435
unknownFullName	**0.450**	0.125	0.421	0.421	0.326	0.413
All Features	**0.817**	**0.752**	**0.811**	**0.794**	0.509	**0.769**

Table 4.11: Efficacy of adding feature groups to the baseline features, measured by MRR, test on Avocado.

Feature Groups	Train/Test on Avocado			Train on Elsayed, Test on Avocado		
	Non-NIL	NIL	Overall	Non-NIL	NIL	Overall
Baseline	0.688	0.171	0.549	0.659	0.190	0.533
ParticipantsGroup	**0.897**	**0.279**	**0.730**	**0.871**	**0.297**	**0.716**
ContactFrequencyGroup	**0.831**	**0.303**	**0.689**	**0.816**	**0.538**	**0.741**
CommunicationCohortGroup	**0.833**	**0.322**	**0.695**	**0.812**	**0.269**	**0.666**
OrganizationGroup	**0.734**	**0.207**	**0.592**	0.669	0.164	0.533
privateContext	0.711	**0.234**	0.582	**0.695**	0.143	0.546
entertainContext	**0.743**	**0.512**	**0.681**	**0.691**	**0.394**	**0.611**
unknownSender	**0.774**	0.134	**0.601**	**0.707**	**0.542**	**0.663**
unknownFullName	**0.739**	**0.753**	**0.743**	**0.730**	**0.481**	**0.663**
All Features	**0.905**	**0.802**	**0.877**	**0.898**	**0.360**	**0.753**

Chapter 5: Meeting Linking for Email

Event linking is a challenging task. In an event, there could be multiple entities involved; the relationships between the entities could be changed by the event; there could be other associated attributes for the event (e.g., time, location). Thus, event linking is a more challenging next step. In this thesis, one particular type of events – meetings – are studied for emails.[1] The task is defined as linking meeting-related email messages to the referenced meetings in a collection-specific meeting knowledge base. The meeting linking system is similar to the entity linking system. The system framework is introduced in Section 5.1, followed by the system design in Section 5.2. The evaluation of the meeting linking system is in Section 5.3. Section 5.4 concludes the chapter.

5.1 System Framework

The framework of the meeting linking system is similar but slightly different from the entity linking system. There are also five stages in the framework for the system, as shown in Figure 5.1: collection-specific meeting knowledge base popula-

[1]This work has been published in Gao, N., Dredze, M., & Oard, D. (2018, January). Enhancing Scientific Collaboration Through Knowledge Base Population and Linking for Meetings. In Proceedings of the 51st Hawaii International Conference on System Sciences [45].

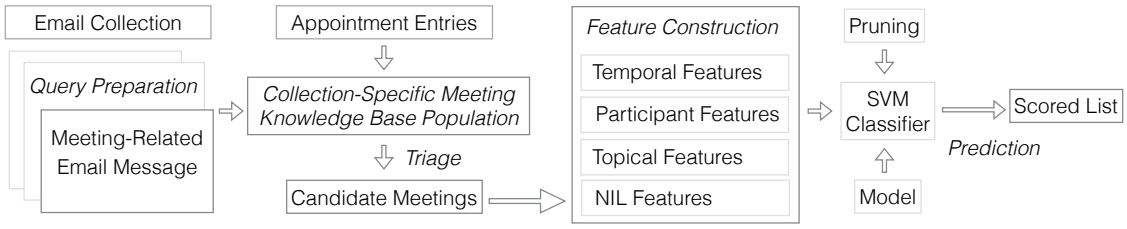

Figure 5.1: System framework for meeting linking.

tion, 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.

The second step is query preparation. The system filters the email collection and selects the email messages that contain the string "meet" in either subject or body of the message. Manual annotation (by the author) of 300 randomly selected email messages 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 an additional 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 referent. NIL is included as a candidate in every case so that the system has the opportunity to rank NIL along with every other candidate.

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. These features are categorized into four groups for presentation purposes. The Support Vector Machine (SVM) regression model nu-SVR from LibSVM [14] 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.

5.2 System Design

There are two stages designed for eliminating the candidates and predicting the true referents for the meeting-related email messages: candidate triage (Section

5.2.1) and ranking (Section 5.2.2).

5.2.1 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, the candidates are selected 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"). The Stanford Temporal Tagger [13] is first used 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 manually created that suggest that the meeting should happen after the message sent date (i.e., agenda, plan, postpone, move) or any of the 5 words on a manually created list that suggest that the meeting occurred before the sent date (i.e., feedback, minutes, notes, recap, report), all candidates in

a 7-day range on that side of the message are retrieved. 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 the list of candidates is further narrowed 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. The calendars of the email message participants is first checked. 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 at least 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, the capitalized words are extracted ("Marketing" and "Group" from the email message subject or the phrase "Marketing Group Meeting") from the subject field of the message and the phrases in the email containing the word "meet" (the phrases are segmented by stop words). The words that containing the string from a manually selected word list (meet, next, today, tomorrow, FW, RE, please, thanks, sorry, nice, great, weekly, minutes, update, request, feedback, agenda, need) are not considered as evidence. Also the words indicating time (e.g., Dec., Wednesday, January) or status (Updated, Cancelled, Accepted) are not considered as evidence supporting retention. 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.

5.2.2 Linking: Ranking

Let $\mathcal{Y}$ be the email messages in the evaluation set (all of which contain the string "meet"), and K_m be the collection-specific meeting knowledge base. For each email message $y \in \mathcal{Y}$ and meeting $m \in K_m$, the system first identifies their extended contexts as $f(E,\ L,\ B,\ U)$, where E represents the participants (sender and recipients) for message y or the owners of meeting m, L is the subject field for y or the meeting subject for m, B is the set of sentences in the email message body that contain the word "meet" for y or the description field of meeting m, and U is the sent date for y or the meeting date for m. Let $\mathcal{M}_i \in K_m$ be the set of candidate meetings for query y_i retrieved from the knowledge base K_m, after triage. Then 18 features $\mathbb{D} = \{\mathcal{D}\,(y_i,\ m)\}$ are computed, where each feature $\mathcal{D}\,(y_i, m)$ is expected to have some predictive value for whether a candidate meeting $m \in \mathcal{M}_i$ is the true referent of the meeting mentioned in email message $y_i \in \mathcal{Y}$. The features are organized here for presentation purposes into four feature groups by the type of evidence that was used for feature construction.

Temporal Features. This set of 2 features is built based on the temporal information of email message y_i and the candidate meeting m. The first feature calculates

the unsigned number of days from the email sent date to the meeting date:

$$\mathscr{D}\left(y_i, m\right) := \mid U_i - U_m \mid. \tag{5.1}$$

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 3.1). Therefore a second feature is built to calculate the minimum absolute days from the meeting date to any of the extracted dates in the email message.

Participant Features. There are 6 features constructed from the participants in the email message. One feature calculates the number of common participants between email message y_i and candidate meeting m:

$$\mathscr{D}\left(y_i, m\right) := \mid \{E_i \cap E_m\} \mid. \tag{5.2}$$

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 for each participant $e \in E_i$ in message y_i. Let $N = \{n\}$ be the known name variants for e. We build one feature to calculate number of participants that have any name variant match in the meeting subject

$$\mathscr{D}\left(y_i, m\right) := \sum_{e \in E_i} |\{n \in N : n \cap L_m \neq\}|, \tag{5.3}$$

Another feature is built to calculate the number of participants that have any name match in the meeting description by substituting B_m for L_m in equation 5.3). Finally, 2 Boolean features are built indicating if there is any name variant match in either the meeting subject or the meeting description:

$$\mathscr{D}\left(y_i, m\right) := \sum_{e\ \in E_i} I(|\{n \in N : n \cap L_m \neq\}| > 0), \tag{5.4}$$

where I is the Indicator function.

Topical Features. Terms indicating the topic of the meeting are extracted from the email message in the triage step. A set of 4 features are built based on the term match between email message y_i and candidate m. For each message y_i, let $\mathscr{K}_i = \{k\}$ be the topic indicative terms. Features are built to calculate the sum of the term frequencies of these terms in the meeting subject L_m:

$$\mathscr{D}\left(y_i, m\right) := \sum_{k\ \in \mathscr{K}_i} TF(k, L_m), \tag{5.5}$$

where $TF(k, L_m)$ is the frequency of term k in meeting subject L_m, or the sum of the term frequencies of the topic indicative terms in the meeting description (substituting B_m for L_m in equation 5.5). 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:

$$\mathscr{D}\left(y_i, m\right) := \sum_{k\ \in \mathscr{K}_i} TF\left(k, L_m\right) * IDF(k), \tag{5.6}$$

where the Inverse Document Frequency (IDF) of term k is calculated based on the union of the subject and description fields of each meeting in the knowledge base, defined in equation 5.7. The description field feature is computed by substituting the use of B_m for L_m in equation 5.6. In general, the more meeting entries the keyword appears in, the less informative it is.

$$IDF\left(k\right) := \log \frac{1}{\sum_{m \in K_m} |k \in \{L_m,\ B_m\}|} * |\mathscr{K}_m| \tag{5.7}$$

NIL Features. There are 6 features constructed to indicate whether the true referenced meeting might be absent from the knowledge base – the NIL case. There is one feature to indicate if the current candidate is the special NIL candidate that added to each list (this allows the ranker to learn to treat the NIL candidate differently if that turns out to be helpful). Other features include: *nilDate* to indicate if there is a specific meeting date in the query email message y_i and there are no candidate meetings on that date; *cancelTermSubject* if there is a term (cancel, n't, not, move, miss) indicating the cancellation of the meeting in the message subject L_i; *cancelTermContext* to indicate if there is one of those same words indicating the cancellation of the meeting in the topical context B_i; *nilIndicative* to indicate if there is no topic indicative term match in any of the candidates; and *cancelStatus* to indicate if the current candidate meeting m is cancelled (with status "Cancelled").

Table 5.1: Statistics on the training and test sets.

	Training	Testing
Meeting-related email messages	4,116	7,276
Meeting entries	7,101	7,254
Total annotations	617	542
Non-NIL annotations	200	160

5.3 Experiments

This section introduces the test collection (Section 5.3.1), followed by the efficacy of linking to known (i.e., Non-NIL) meetings (Section 5.3.2), separately analyzes the utility of each feature group (Section 5.3.3), and conducts a feature addition study (Section 5.3.4). Finally, the linking for NIL cases is discussed (Section 5.3.5).

5.3.1 Test Collection

To evaluate the efficacy of the proposed meeting linking system, the email collection and the meeting knowledge base are split into disjoint training and testing sets. The 226 email accounts with appointment entries are randomly divided into the training and test sets of equal size. The training set includes the 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 test 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 test is constructed solely from the calendars

Table 5.2: Effectiveness measures, Non-NIL queries.

	Accuracy	MRR
Random	0.312	0.501
Our system	0.899	0.930

of the test accounts.

Table 5.1 shows the basic statistics on the training and test sets. The author of this thesis annotated 617 randomly selected meeting-related email messages (Total annotations) and was able to link 200 messages (Non-NIL annotations) in the training set to the meeting entries. For the remaining 417 email messages, the author 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 160 Non-NIL annotations are used to evaluate the efficacy of our system on linking email messages to the referenced meeting entries. The analysis for the system predictions on the NIL links is shown in Section 5.3.5.

5.3.2 Linking for Non-NIL

Table 5.2 shows the efficacy of linking Non-NIL query email messages to the referenced meeting entries. Since there is no prior work on the same task, randomly ranking the triaged candidate meetings for a query message is defined as a low baseline. The Accuracy for the Baseline is 0.312, which reflects the sharply skewed distribution of triage results. The triage step (Section 5.2.1) 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 20.6% of the cases account for 0.206 of the 0.312 observed accuracy. The proposed system is able to nearly triple that Baseline accuracy by using all the features (Section 5.2.2). In the next sections, the efficacy of each feature group individually (Section 5.3.3) and in combination (Section 5.3.4) are explicitly analyzed.

5.3.3 Single Feature Groups

Figure 5.2 shows the MRR 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 only features in that feature group. The Accuracy for Random and All (using all four groups of features) are also shown in Figure 5.2 for reference. Topical features are the best single feature group (0.78 MRR), 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

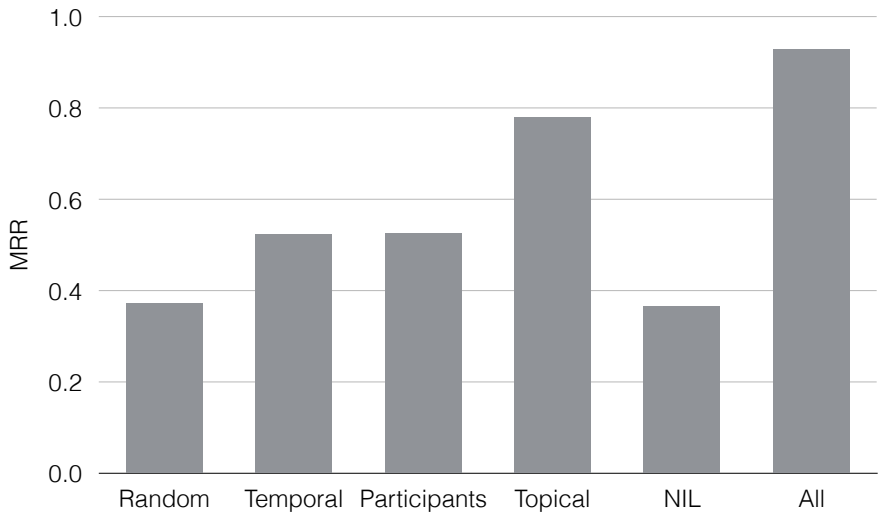

Figure 5.2: MRR for each single feature group.

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.

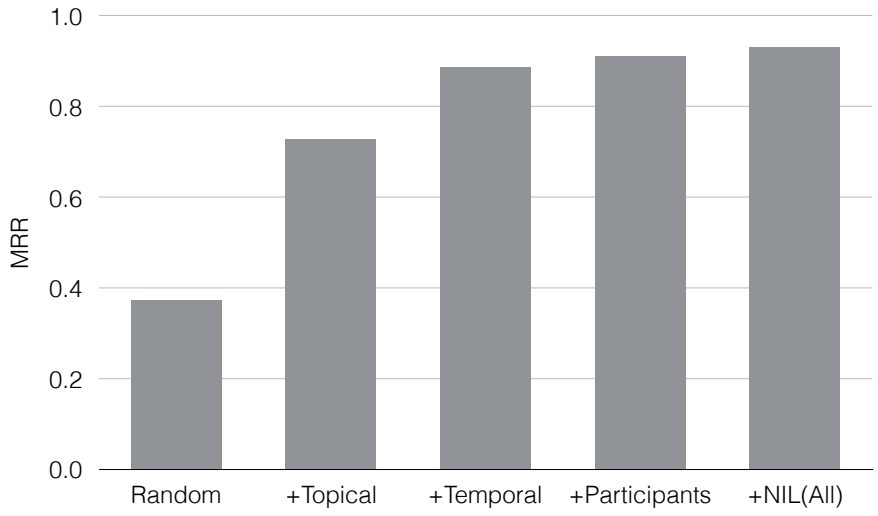

Figure 5.3: Feature group addition.

5.3.4 Feature Group Addition

Figure 5.2 shows that none of the single feature groups achieves an MRR near that of the full set of features. MRR thus benefits from the combination of complementary evidence captured by different feature groups. Figure 5.3 shows the results of cumulatively adding feature groups. From left to right, the Baseline is again the random selection case in which no ranking features are used. Then the feature group that provides the greatest gain in the MRR (Topical) is added, yielding an MRR of 0.727. Next each remaining feature set is added, finding that the combination of Topical and Temporal features achieves the highest MRR (0.883). This is close to the result for using all feature groups (0.930). Small improvements result from further adding the most helpful of the two remaining feature sets (Participants) and then from adding NIL features.

5.3.5 Linking for NIL

In the test set, the 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 Elsayed's set of NIL annotations for person entity linking in email. 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, NIL cases are artificially created 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 the same process is done in the training set, thus training on 10% fewer Non-NIL cases) and the MRR on those 16 (i.e., 10% of 160) artificially created

true NIL queries in the testing set is 0.512.

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 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. Lacking better candidates, however, our system makes a prediction, albeit with low confidence. While making the annotations, the human annotator is provided with full information on both meeting entries (the correct one and the wrong one), while with our NIL simulation design, the system is only provided with part of the information (the wrong meeting entry). This also invalidates the design of some of the NIL features such as *nilDate* and *nilIndicative*. *nilDate* feature recognizes NIL references when there is no candidate on the specific meeting date. However, in our simulation case, the referent is NIL while there is a candidate on the specific meeting date. *nilIndicative* recognizes NIL referents when there is no candidate with topic indicative term match. However, in the simulation

case, the reference is NIL while there is a candidate with topic indicative term match. Because of the unequal input information to the human annotator and the system, and the invalidate of the NIL features, the results on NIL simulation might yield overly pessimistic results as an indicator of what could be achieved in practice.

5.4 Chapter Summary

This chapter introduced a meeting linking system links the meeting-related email messages to the collection-specific meeting KB. The meeting entries can be easily recognized and extracted with high precision and recall from the calendars by following manually designed rules. The system is built following the construction of the entity linking system. Different triage steps and features are designed to adjust the new meeting linking task. Different from the entity linking system, the triage step is particularly important for the meeting linking task. Considering a named mention "John" for the entity linking task, all the person entities named John should be considered as referenced candidates. For the meeting linking task by contrast, the query is an email message rather than a named mention, which provides much richer information such as the time period of the meeting or the topic of the meeting. A well designed triage step could thus spontaneously retrieve possible candidates (i.e., being recall-oriented) and eliminate meeting entries that are not the true referent (i.e., being precision-oriented) at the same time. By substantially narrowing down the number of candidates, the system provides less noisy candidates to the next ranking step, and thus improves the efficacy of the system.

The meeting linking system works quite well when the mentioned meeting is present in the knowledge base, although the present implementation is a tad overeager to make a link when none should be made. An imperfect simulation of NIL cases is also conducted. For the errors the system made on NIL cases, there are two cases: misleading evidence and low evidence. For the misleading evidence cases, the system tends to make the same mistakes as the human annotators. For the low evidence cases, the efficacy of the system might be improved by more training data and improved feature design. In the future work, a better designed test collection is needed for the NIL annotations. The human annotators perhaps might provide more information on the reasons and also the confidence level of the NIL decisions.

The proposed meeting linking system is motivated in part by the application of scientific collaboration, as discussed in Section1.3.3. By using the proposed meeting linking system, the email messages related to a meeting entry could be automatically linked. The practical use of the system requires high accuracy for both Non-NIL and NIL cases. The accuracy of the Non-NIL links guarantees that when there are email messages that are related to the meeting, they will be linked and presented to the user. The relatively lower NIL accuracy indicates that there will be more false positive links from email messages to the meeting entries.

Chapter 6: Entity Linking for Conversational Speech

The knowledge base population related tasks have been well studied for dissemination-oriented sources as introduced in Section 2.1. However, considering the information generated each day, vast majority of them are conversational – the amount of language produced by a person on daily basis is 16,000 words on average [95]. However, there are few studies on the task of entity linking for conversational speech (as reviewed in Section 2.6). In this thesis, the Enron phone recording collection is used as a conversational speech dataset. Section 6.1 discusses the linking of three types of named mentions (person, organization, location) detected in the manual transcripts of phone records to the general knowledge base Wikipedia and the collection-specific knowledge bases built from the associated Enron email collection.

The efficacy of entity linking for named mentions of people on phone recordings benefits substantially from the recognition of the speakers and the social network between all the speakers in the collection. Speaker identification techniques could be used to automatically recognize the speakers for the phone recordings, but the recording quality, informal speak styles, and the background noise conversely affect the speaker identification efficacy for conversational content. Section 6.2 explores the use of side information to improve the efficacy of the speaker identification.

Experiments also show a positive effect on the person entity linking task for Enron phone recordings. These results are presented in Section 6.3.

6.1 Entity Linking for Conversational Speech

This section first introduces the test collection built from the Enron Phone Recordings in Section 6.1.1. Then Section 6.1.2 discusses the named entity recognition task for conversational speech, followed by the evaluation of the entity linking system (introduced in Section 4.2) on the task of linking named mentions of three types (PER, ORG, LOC) to the general knowledge base Wikipedia, and collection-specific person and organization knowledge bases in Section 6.1.3.[1]

To apply the entity linking system on the phone recordings, there are several tweaks on the features: (1) all the speakers are treated equally as conversation participants. Therefore, the features *inHeaderExceptFromField*, *inToField*, *inFromField*, *inCcField*, *inBccField*, *inToFieldNormalized*, *inCcFieldNormalizedinFromContacts* and *inBccFieldNormalized* are set as 0; (2) the **TopicalContextGroup** features are calculated based on manual transcripts of the Enron phone recordings; (3) there is no difference between the conversation and the conversation thread. Therefore, all the features based on "thread" (*inThreadHeader*, *candidateInThread*, *fuzzyCandidateInThread*, *candidateInThreadProb*, *fuzzyCandidateInThreadProb*, *inContactsThread* and *inContactsThreadNormalized*) are set as 0.

[1]This work has been published in Gao, N., Oard, D. W., & Dredze, M. (2017, August). Support for interactive identification of mentioned entities in conversational speech. In International Conference on Research and Development in Information Retrieval (SIGIR) (pp. 953-956) [49].

Table 6.1: Human annotations for the linking.

	All	PER KB	ORG KB	Wikipedia	NIL
PER	279	260	0	12	15
ORG	174	0	142	81	32
LOC	96	0	0	75	21
Total	549	260	142	168	68

6.1.1 Test Collection

The author of this thesis annotated named mentions and KB links (including NIL) in all three KBs (Wikipedia, collection-specific person KB, collection-specific organization KB) for the 540 PER, ORG and LOC mentions in the 64 manual transcripts.[2] For PER and ORG mentions, a referent entity might be present in both the Wikipedia KB and the corresponding collection-specific KB (e.g., Enron). The mentions include misspellings (e.g., Holli misspelled as Holly), abbreviations (e.g., LV Co-gen), and initials (e.g., ISO). Most of the person mentions are first names or nicknames (e.g., Ken). Table 6.1 summarizes the linking annotations used as ground truth. A second annotator independently linked a randomly selected half of the PER name mentions. This yielded an exact match agreement of 0.78 for the cases in which the author had made a link. A third annotator independently linked 20 randomly selected ORG and 20 randomly selected LOC mentions. The agreement with the author on the ORG and LOC mentions is 0.85 and 0.90, respectively.

[2] `http://www.umiacs.umd.edu/~ninggao/publications`

6.1.2 Named Entity Recognition

As discussed in Section 2.1, successful named entity recognition (NER) is a prerequisite to entity linking in the pipeline of knowledge base population. Systems designed for the NER task (e.g., the Illinois Named Entity Tagger used in Section 3.2) achieve high precision and recall for dissemination-oriented text and speech. However, the system performance on the conversation speech collections is questionable due to the poor quality of ASR. In this section, we discuss the performance of entity detection for Enron Phone Recordings. The Stanford NER system [39] is used to automatically identify person mentions (we refer to this as "Auto"). Two different training sets are considered for the NER tagger: (1) training on text only data from CONLL (2003 [129], or 2008 [126]); (2) training on both text and speech by adding in data from ACE 2005.[3] This approach works rather well for manual transcripts, but with Automatic Speech Recognition (ASR) transcripts (or the quality that we have available) it exhibits serious deficiencies.

We begin by evaluating the accuracy of NER using the Stanford NER system trained on text (T) or text and speech data (T+S). Results are reported for both the manual and ASR transcripts. Table 6.2 reports precision, recall and F_1 for detection of entity mentions on *Manual* and *ASR* transcripts on the test set. Even with text-only training, Table 6.2 shows that NER performance for this manually transcribed conversational telephone speech is already comparable to what we would expect as state-of-the-art performance for NER on newswire text, and the addition of speech

[3] `https://www.ldc.upenn.edu/collaborations/past-projects/ace`

training data yields a slight further improvement in recall.

In contrast, NER performance on ASR output is markedly worse, although again the additional speech training data yields small improvements (in both precision and recall). Table 6.3 provides an error analysis. For *Manual* transcripts, all the entity detection errors are (of course) caused by the NER system. For *ASR* transcripts, the detection errors could be caused by the errors in ASR transcripts or the NER system. For each misrecognized mention, if there is a (case-insensitive) exact string match between the query mention and any point in the ASR transcript, we code the error as an NER error; otherwise, we code it as an ASR error. As Table 6.3 shows, the dominant cause of NER failures on ASR is ASR errors in which the mentioned name is simply not correctly transcribed and thus could not have been found by NER. In future work it may be possible to improve the overall NER results by using spoken term detection techniques to detect the presence of specific names that are of interest, even when those names are out-of-vocabulary for the ASR system.

6.1.3 Experiments for Entity Linking

Section 6.1.3.1 shows the experiment results of entity linking for all three types. Section 6.1.3.2 explicitly discusses the linking of person named mentions to the collection-specific person knowledge base built from Enron email collection on both manual and ASR transcripts. The knowledge bases used for the Enron phone recordings are identical to the knowledge bases for the Enron emails 4.5.

Transcription		Manual		
Evaluation Measure		P	R	F_1
PER	T	0.881	**0.910**	0.899
	T+S	**0.913**	0.905	**0.909**
ORG	T	0.562	0.621	0.590
	T+S	**0.566**	**0.627**	**0.595**
GPE	T	0.624	0.692	0.656
	T+S	**0.640**	**0.709**	**0.673**
Transcription		ASR		
Evaluation Measure		P	R	F_1
PER	T	0.173	0.210	0.190
	T+S	**0.174**	**0.227**	**0.197**
ORG	T	**0.079**	0.077	0.078
	T+S	0.078	**0.085**	**0.081**
GPE	T	**0.144**	0.337	0.201
	T+S	**0.144**	**0.340**	**0.202**

Table 6.2: Named entity recognition on Manual or ASR transcripts; trained on Text (T), or Text and Spoken (T+S) language; measured by Precision (P), Recall (R) and F_1.

6.1.3.1 Entity Linking for Three Types

The entity linking system is evaluated on all three mention types. Table 6.4 shows the MRR of the system using *All features* on the named mentions extracted from the manual transcripts. Two baselines are constructed : (1) only the *General* feature group, and (2) a *Random* baseline that randomly selects one entity from the triaged candidate set. As expected, the system does much better than the baselines for all three entity types when evaluated on All mentions in the test set.

Using only *General* features, the linking result for PER is much worse than for ORG or LOC. Unlike in news articles, mentions of people in these conversations are mostly just first names (e.g., John) or nicknames (e.g., Bill), many of which result in hundreds of candidates (an average of 314 in the collection). Moreover, as shown

		Manual	ASR
PER	Correct	253	63
	ASR Error	0	169
	NER Error	26	47
ORG	Correct	109	15
	ASR Error	0	133
	NER Error	65	26
GPE	Correct	68	33
	ASR Error	0	41
	NER Error	28	22
All	Correct	430	111
	ASR Error	0	343
	NER Error	119	95

Table 6.3: Count of Correct NER results and NER failures due to ASR missing the mentioned name (ASR error) or due to some other NER error.

Table 6.4: Entity linking for all mentions.

		Non-NIL	NIL	All
PER	Random	0.055	0.167	0.060
	General	0.253	0.612	0.273
	All features	**0.786**	**0.669**	**0.779**
ORG	Random	0.243	0.32	0.301
	General	0.498	**0.821**	0.557
	All features	**0.843**	0.612	**0.800**
LOC	Random	0.184	0.200	0.188
	General	0.451	**0.567**	0.476
	All features	**0.811**	0.474	**0.737**
All	Random	0.134	0.371	0.164
	General	0.356	**0.695**	0.466
	All features	**0.807**	0.583	**0.776**

in Table 6.1, most (90%) of the named mentions of people refer to entities that can only be found in the collection-specific PER KB, and that KB contains less and sparser context than the Wikipedia KB. Human disambiguation of entity mentions in conversational speech relies heavily on shared context, and indeed we observe that by adding *All* features, the MRR for PER mentions improves from 0.273 to 0.779. The entity linking system shows similar efficiency on emails and phone recordings,

comparing 0.748 to 0.779 for PER, 0.812 to 0.800 for ORG, and 0.800 to 0.776 for LOC.

Many of the errors in linking ORG mentions arise from changes in organization names due to mergers and acquisitions, which change the name of a company. For example, Reliant Energy (one of the ORG mentions) was renamed NRG Energy after the conversation was recorded, but before the construction of the ORG KB. Additional information (e.g., the Wikipedia edit log) might help to resolve such errors. For LOC mentions, lack of context in the conversation is the main reason for the errors. For example, the speakers mention "Four Corners" in a short conversation without specifying the US state. Without additional context, it is difficult to know if the location is "Four Corners, California" or "Four Corners, Oregon". This problem could potentially be solved if there were other conversations between the same group of speakers available. For example, the same speakers mentioning "Four Corners" together with "California" in a recent conversation might indicate the referent to be "Four Corners, California".

It is (on average) harder for the system to correctly detect NIL references that should not be linked than it is to link Non-NIL references to the correct entity. Considering both NIL and Non-NIL references, the overall MRR for each entity type is in a fairly narrow range between about 0.7 and 0.8, indicating that the correct referent (or NIL) is often found in the first or second position in the ranked list. These results are below the scores typically reported for newswire (for which MRR above 0.9 is commonly reported), but with feature designs that model some of the context available to the participants, the system can achieve linking accuracy that

could be useful.

6.1.3.2 Linking Person Mentions

Similar to the entity linking task for email, the linking of person named mentions is particularly interesting for phone recordings due to its significantly larger ambiguity and lack of context. The efficacy of the proposed entity linking system for person named mentions is evaluated on both manual and ASR results. For both cases, the person named mentions are provided to the entity linking system as queries. There are two baselines:

- **Random,** randomly order the list of candidates returned by triage.
- **Contact Frequency,** order the list of candidates by the frequency with which each has contacted the speakers in the call (with ties broken randomly). There are two contact frequency baselines, one for the phone call communication graph and one for the email communication graph.

Table 6.5 reports results for the baselines, entity linking models trained for each feature set individually, and all features together. The results for features computed based on the email communication graph and the phone communication graph are reported separately. There are a few observations about these results. First, the results on manual and ASR transcripts are comparable when using manually recognized named mentions. This indicates that the context of the mention, which could be corrupted by poor ASR, does not substantially impact linking accuracy. This observation suggests that manually designated mention queries are

practical, even when the system only has access to ASR transcripts. Second, the social context features provide a substantial boost in performance; it thus seems that leveraging the shared context of the speakers even in this limited way is very useful. Third, it is interesting that we see larger MRR gain when features are extracted from the phone call communication network than when the same features are extracted from the email network, even though there is much more data in the considerably larger email communication network. One possible explanation for this is a "small world" phenomenon: the phone call network involves a far smaller number of participants, many of whom have participated in a substantial number of phone calls. This suggests that there likely are behavioral differences between the two networks: perhaps speakers talk on the phone about other people who they tend to call rather than email. Nevertheless, combining features estimated separately on each communication graph still yields somewhat better results than using either network alone.

One potential concern that might arise with the results in Table 6.5 is that named mentions of participants in the same call (e.g., "Hi this is Bill ...") might be inflating the averaged results. Correctly linking mentions to participants is sometimes important (as with "Bill, did you say you would do that?" on a multi-party call), but it would expected that such mentions to be easier to resolve because the identities of known participants are provided to the system as features. Table 6.6 shows results similar in structure to those in Table 6.5, but with the 99 participant mentions removed from the mention queries, leaving 170 **non-participant** mentions for the evaluation. As can be seen, this change reduces the MRR somewhat

Transcription	Manual	ASR
Random	0.162	0.162
General	0.626	0.626
Contact Freq (P)	0.288	0.288
Social (P)	0.613	0.613
Topical (P)	0.590	0.568
All features (P)	0.763	0.754
Contact Freq (E)	0.245	0.245
Social (E)	0.283	0.283
Topical (E)	0.261	0.257
All features (E)	0.649	0.634
Social (P+E)	0.653	0.653
Topical (P+E)	0.591	0.572
All features (P+E)	**0.753**	**0.713**

Table 6.5: Entity linking using context extracted from phone recordings (P) or emails (E), measured by MRR.

for every condition, but the MRR is still above 0.5 if either manual transcripts or manual NER are available.

Another potential concern that could arise is that some mentions are naturally easier to resolve than others. In particular, a full-name mention (e.g., Bill Clinton) will naturally be much less ambiguous than a first-name mention (e.g., "Bill", or even "Clinton"). Therefore the test collection is further ablated to remove all multi-token named mentions of nonparticipants. Table 6.7 shows MRR results averaged over the remaining single-token nonparticipant mentions. As can be seen, the adverse effect of this restriction to single-token mentioned is small, for example reducing the MRR on ASR from 0.521 to 0.504.

Transcription	Manual	ASR
Random	0.083	0.054
General	0.233	0.208
Contact Freq (P)	0.154	0.148
Social (P)	0.443	0.392
Topical (P)	0.507	0.323
All features (P)	0.541	0.502
Contact Freq (E)	0.032	0.027
Social (E)	0.290	0.277
Topical (E)	0.238	0.221
All features (E)	0.397	0.328
Social (P+E)	0.457	0.429
Topical (P+E)	0.509	0.394
All features (P+E)	**0.561**	**0.521**

Table 6.6: Entity linking only for mentions that refer to **nonparticipants**; context extracted from phone recordings (P) or emails (E), measured by MRR.

Transcription	Manual	ASR
Random	0.054	0.041
General	0.144	0.107
Contact Freq (P)	0.080	0.068
Social (P)	0.400	0.356
Topical (P)	0.459	0.322
All features (P)	0.530	0.493
Contact Freq (E)	0.027	0.019
Social (E)	0.245	0.211
Topical (E)	0.196	0.175
All features (E)	0.346	0.302
Social (P+E)	0.432	0.391
Topical (P+E)	0.460	0.324
All features (P+E)	**0.541**	**0.504**

Table 6.7: Entity linking only for **single token** mentions that refer to **nonparticipants** using context extracted from phone recordings (P) or emails (E), measured by MRR.

6.2 Speaker Identification

The experimental results in Section 6.1.3.2 show that the entity linking results rely largely on the recognition of recorded speakers and the social network between all the candidate speakers. However, only a small fraction of the recordings in the

collection used have manually recognized speakers. Thus for the vast majority of the recordings, the effectiveness of the entity linking system is unsatisfactory. Speaker identification is the task of automatically recognizing the speakers in audio files from a collection of speaker candidates. This could be used to improve the entity linking performance on the recordings with unknown speakers. However, most existing speaker identification systems leverage only acoustic evidence. In this section, the use of side information to improve speaker identification is also explored [4].

The test collection built from the Enron phone recordings is first introduced in Section 6.2.1. Section 6.2.2 then introduces the evaluation measures. Baseline results using acoustic evidence alone are presented in Section 6.2.3, followed by results using five types of contextual features in Section 6.2.4. The discussions of those results are presented in Section 6.2.5.

6.2.1 Test Collection

Of the 64 manual transcripts, only 57 can be matched with telephone recordings. Thus only these 57 recordings with true speakers recognized in the manual transcripts are used in this section. These 57 recordings are partitioned into a training set containing 28 recordings and a test set containing 29 recordings. Across the 57 recordings there are a total of 41 different speakers whose names are available from the transcripts, and the author is able to manually associate 37 of these names with the full names of people represented in Elasyed's collection-specific knowledge

[4]This work has been published in Gao, N., Sell, G., Oard, D. W., & Dredze, M.. Leveraging side information for speaker identification with the Enron conversational telephone speech collection. In Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE (pp. 577-583) [50].

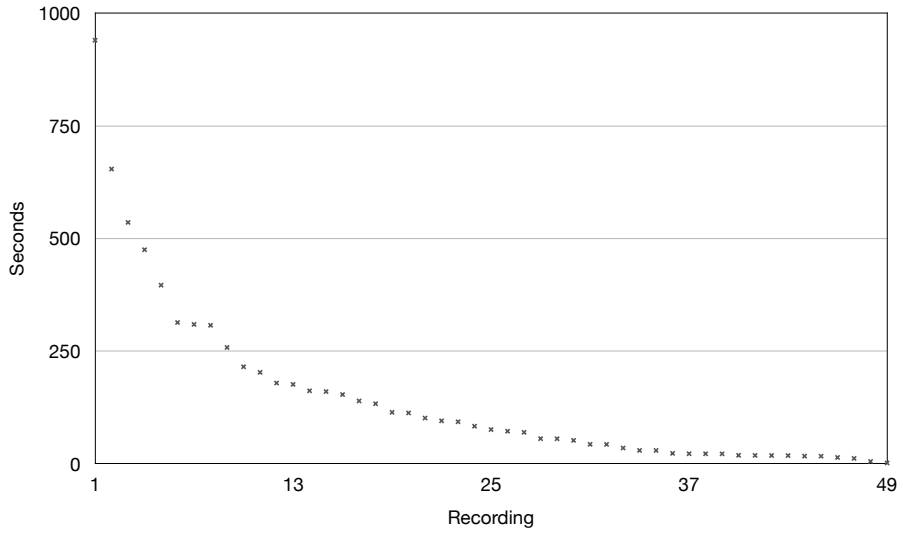

Figure 6.1: Duration of the testing audio files.

base. 28 of the 57 recordings are selected as a training set in a manner that ensured that all 41 known speakers would each be represented in at least one training recording. The 28 recordings in the training set are then manually diarized into short segments for each known speaker for use in training the speaker models. The remaining 29 transcribed recordings are then used as the basis for the test set. The test set is manually segmented into individual telephone calls; this results in a total of 49 test calls. Some of the audio files are very short. Figure 6.1 shows the duration of the 49 test set files. Each dot in Figure 6.1 represents an audio file, shown in descending order of their duration. The known speakers for each call were manually determined from the scanned transcripts, which had been manually prepared.

6.2.2 Evaluation Metric

The core task of this section is to identify which speakers from the training set are present in a call from the test set. There is always at least one known speaker, and often there are two. There are some calls with three or more speakers, but none include more than two known speakers. To get an insight into identification performance, an evaluation measure is supplemented based on mean reciprocal rank (MRR), common in information retrieval. The fundamental statistic that this section seeks to estimate is the rank of each known speaker in the list of scores for a particular recording. When there are multiple known speakers, the system's ranked list is replicated with one of the two known speakers removed. Then the metric is computed based on the harmonic mean of the rank of the known speaker in each such list:

$$R = \frac{|R|}{\sum_{i=1}^{|} Q| \frac{1}{r_i}} - 1, \tag{6.1}$$

where r_i is the rank of the known speaker in list i and $|Q|$ is the number of lists (i.e., the number of speaker-call pairs). R can thus be interpreted as the number of rank positions below what would otherwise be a perfect ranking at which the system places the correct speaker. In the experiments R is always between zero and one, but in principle R is unbounded and R for a random ranking would be about 20.

Lower values are are preferred, with zero being the lowest possible value. Moreover, because zero indicates perfect performance (corresponding to consistently putting every known speaker at the earliest possible rank), R is a ratio measure (i.e.,

a measure in which, for example, a value twice as large is twice as bad); this makes percentage differences meaningful.

6.2.3 Acoustic Speaker Identification

In this section, the text-independent speaker recognition system [119] is first used to get the estimated speakers for all the recording files. Within the 1,731 recording files, 28 files with known speakers are used as the training set; 29 files (in total 49 calls after manual segmentation) with known speakers are used as the test set; the speakers of the remaining 1,664 files are predicted automatically by the system. For each recording file, there are 41 speaker candidates. From the training files, the author of this thesis manually diarized audio samples for each speaker candidate. Then the speaker recognition system calculates the similarity between the vectors representing the acoustic features extracted from the audio files and the speaker samples. The candidates for each file are ranked by their probability of being a true speaker, thus the problem is treated as a closed-set ranking task in this dissertation. The baseline speaker recognition system achieves an R of 0.73. Next some methods to improve upon this baseline with side information are introduced.

6.2.4 Re-Ranking Techniques

In this section, approaches for re-ranking speakers using a social network (Section 6.2.4.1), channel information (Section 6.2.4.2), or name variant detection (Section 6.2.4.3) are introduced.

Acoustic Rank	Ranking of Speaker Pairs	Final Re-ranked List
Speaker01	*Speaker01* & *Speaker03*	*Speaker01*
Speaker02	*Speaker04* & *Speaker01*	*Speaker03*
Speaker03		*Speaker04*
Speaker04		*Speaker02*

Table 6.8: A re-ranking example.

6.2.4.1 Social Network Re-Ranking

Some of the most interesting experiments in this Section involved re-ranking using a social network. The simplest such case is the email social network. There are 41 known speakers, 37 of whom sent or received email in the CMU Enron email collection. For each of these 37 speakers, it is known from Elsayed's knowledge base that how often they communicated with each of the other 36 of the known speakers in the email collection. A conversation to more often involve frequent communicants than rare ones would be expected.

This intuition is formalized as follows. If two known speakers were present in the same email header (i.e., if one sent and the other received an email message, or if both received the same message), there is an edge built between them in the social network, and the weight of that edge is set to be the frequency with which they communicate. Let g_l denote the sum of the edge weights that are connected with one of the speakers (which we refer to as the left speaker), g_r to denote the sum of the edge weights that are connected with the other (right) speaker, g_{lr} to denote the (undirected) edge weight between the left speaker and the right speaker, and $\sum g$ to denote the sum of all the edge weights in the social network. The score

of a pair is then calculated as:

$$s_p = \frac{1}{2}\Big(\Big(1 + \frac{g_l}{\sum g}\Big)s_l + \Big(1 + \frac{g_r}{\sum g}\Big)s_r\Big)\Big(1 + \frac{g_{lr}}{\sum g}\Big). \tag{6.2}$$

As can be seen from the formula, there are five factors that influence the estimation of whether the left and right speakers are true speakers in the conversation: the acoustic score s_l of the left speaker, boosted by the degree to which the left speaker is a frequent communicant ($\frac{g_l}{\sum g}$); the acoustic score s_r of the right speaker, boosted by the degree to which the right speaker is a frequent communicant ($\frac{g_r}{\sum g}$); and a boosting factor applied to both that reflects the degree to which these two speakers communicate with each other ($\frac{g_{lr}}{\sum g}$). The use of two individual boosting factors is a precision-oriented design reflecting that only frequent communicants with high acoustic ranks have the power to "pull" up other speakers. The system then re-ranks the speakers according to their highest associated s_p (or their original score in the case of speakers with no observed pairs). Table 6.8 illustrates the ranking by acoustic score, the pair ranking, and the final re-ranked list using an actual example from the collection (with names anonymized). The first pair places *Speaker01* and *Speaker03* on the re-ranked list, in that order; the second pair then results in addition of *Speaker04*, and the final insertion of speakers missing from any pair adds *Speaker02*.

If it were known which speakers had actually participated in each call in some large set of phone calls, the system could apply a similar process to leverage the telephone social network, but true labels are only known for a small number of phone

calls. Instead, the acoustic baseline system described in Section 6.2.3 is used to predict which speakers participated in each of the 1,703 non-training recordings (1,731 minus the 28 labeled training recordings). By counting these predicted telephone interactions, a similar network can be generated to that drawn from the emails, thus producing an alternative re-ranking that can be evaluated to determine whether the larger size and more accurate observability in the email social network yields better results than the smaller and less accurately estimated, but perhaps more highly comparable, telephone social network. Table 6.10 shows the results. The telephone social network turns out to be the clear winner, improving by 11% relative to the baseline (from 0.73 to 0.65) compared to 0.70 for the email network.

Aggregate results can mask important insights, so Figure 6.2 provides a compact visualization of where this approach works, and where it fails, for the self-trained telephone social network. In this plot, the Y axis shows the change in rank of the true speakers as a result of the side information for each test trial, which are itemized on the x-axis and sorted by initial ranking. The upper and lower bounds of possible rank changes are also shown for context.

As can be seen, no speaker that the acoustic evidence had initially correctly placed at the best possible rank (i.e, no speaker for which the upper bound on the possible improvement was zero) was adversely affected by re-ranking. Notably, four speakers (each of which started out near the top of the list) achieved the maximum possible improvement. Re-ranking resulted in more changes—both positive and negative—for speakers lower in the list, moving the rank up in 17 cases and down in 12.

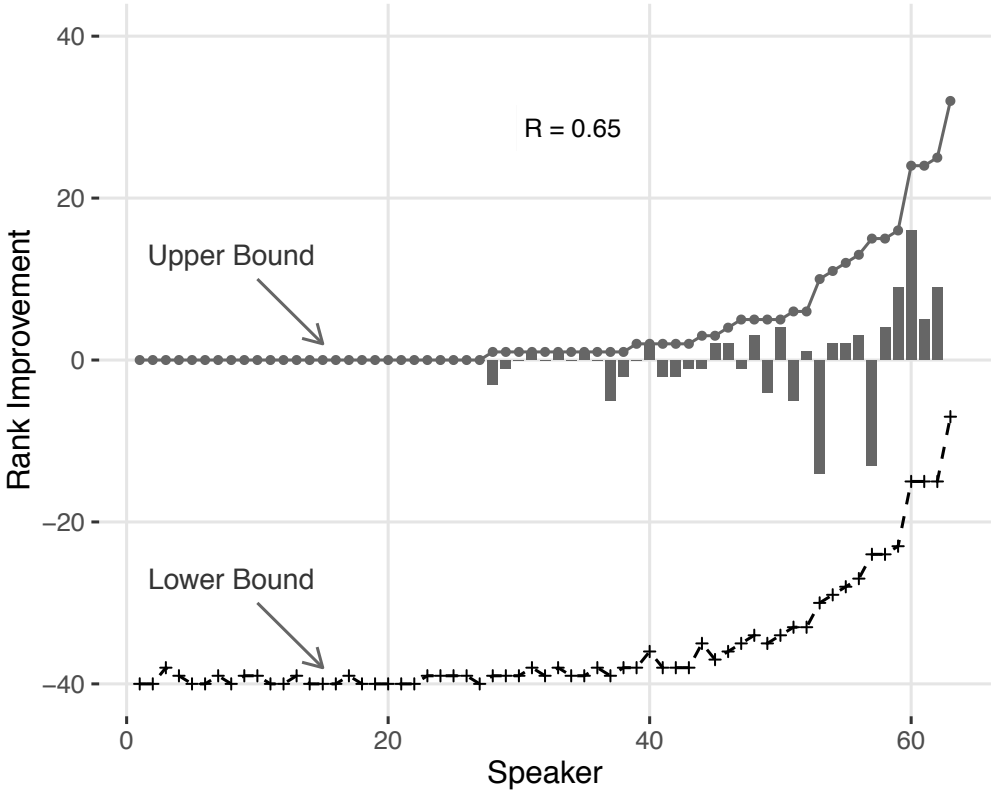

Figure 6.2: Rank improvement of true speaker after re-ranking by **self-trained social network**. X-axis represents the speaker instances. Y-axis shows the change in rank for each speaker instance. For each speaker instance ranked at position *i*, improvement upper bound is *i-1*, and the improvement lower bound is *number of candidates - i*, where *number of candidates* is the number of candidates for each speaker-recording instance with other known speakers removed.

6.2.4.2 Channel Re-Ranking

The Enron Phone Recording collection also includes metadata indicating on which channel each call was made, as well as a list (prepared professionally for use in a lawsuit) that indicates which people were typically recorded on which channels. Table 6.9 shows an anonymized excerpt from this list. The "Main Channel(s)" are those on which the compiler of the list expected to see the speaker most often, whereas "Other Channels" are those or which they chose to note that the speaker

Speaker	Main Channel(s)	Other Channels
Speaker05	1, 13	2, 3, 14, 15
Speaker06	26	16, 25, 51

Table 6.9: Example channel information for speakers.

was also sometimes present. Some channels repeat as main channels for different speakers, suggesting that there was some sharing of phones (e.g., during different work shifts), meaning that this channel information is not sufficient on its own for predicting the true speaker. It is easy to see how the system might use this information to re-rank the speakers, since if it is known that *Speaker05*'s main channel is channel 1 and that channel 1 is not *Speaker06*'s main channel, then *Speaker05* may be a better speaker candidate than *Speaker06* when the call is recorded on channel 1.

For comparison with this manually compiled channel information, a process similar to that used to build the telephone social network is used to estimate channel probabilities for each speaker on the 1,703 non-training recordings. To do this, the observed channel mappings from the training recordings are used to estimate how often each speaker was likely to be recorded on each channel. Then the re-ranking process is formalized as follows. Let $h = (h_1, \ldots, h_m)$ be the m unique channels on which recordings in the collection have been recorded, and $W_c = (w_1, \ldots, w_m)$ be the number of calls in which candidate speaker c was detected using each channel based on acoustic evidence. The system then calculates a new score $s_c^{'}$ for each candidate c based on the acoustic prediction score s_c and the maximum likelihood

estimate of the probability that speaker c is observed on channel i:

$$s_c^{'} = \Big(1 + \frac{\lambda w_i}{\sum_{q=1}^{m} w_q}\Big) s_c, \tag{6.3}$$

where λ is a parameter to adjust for the relative weight of the channel information. In the experiments, λ is arbitrarily set as 1.

To use the same re-ranking process with the manually prepared list, the system arbitrarily sets the number of calls to 2 for main channels, to 1 for other channels, and and 0 for channels that are not listed. Although this process is not optimized, it serves as a useful reference to which the results of the automated estimates that are estimated from a larger, but noisier, set of examples from what amounts to semi-supervised training can be compared. Using the manually prepared table improves R (from 0.73 to 0.53), while using the automatic channel estimates improves R somewhat less (to 0.57). This improvement in R from the fully automated technique is a 22% relative improvement that is significant under a two-tailed paired t-test (at $p < 0.05$).

Figure 6.3 shows a compact analysis of the case-by-case results for the provided speaker-channel table that is structured identically to that in Figure 6.2. In this case there are 11 improvements and 12 reductions in rank, but many of the improvements are near the top of the ranked list and at or near the upper bound, whereas the reductions in rank occur only for correct candidates that were already at or below rank 3, and they come nowhere near the lower bound. R rewards these improvements more than it penalizes reductions in rank with those characteristics.

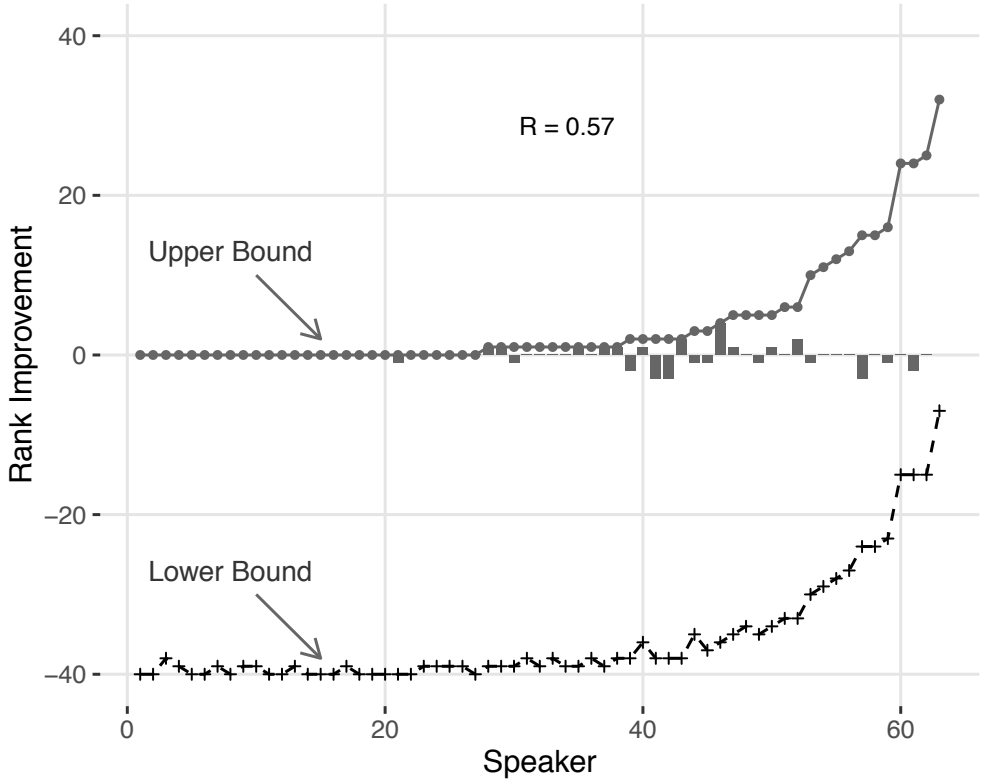

Figure 6.3: Rank improvement using the **self-trained channel** information. X-axis represents the speaker instances. Y-axis shows the change in rank for each speaker instance. For each speaker instance ranked at position *i*, improvement upper bound is *i-1*, and the improvement lower bound is *number of candidates* - *i*, where *number of candidates* is the number of candidates for each speaker-recording instance with other known speakers removed.

6.2.4.3 Name Mention Re-Ranking

Frequently speakers will identify themselves at the beginning of a conversation (e.g., "Snohomish, Jay." "Hey Jay, Holly."). Since it is known (from the knowledge base constructed from the email collection) how 37 of the 41 speakers might be referenced, the evidence from named mentions can easily be used. This is formalized as follows. For each speaker candidate, the system first matches it to at most one person entity in the collection-specific person knowledge base built from the

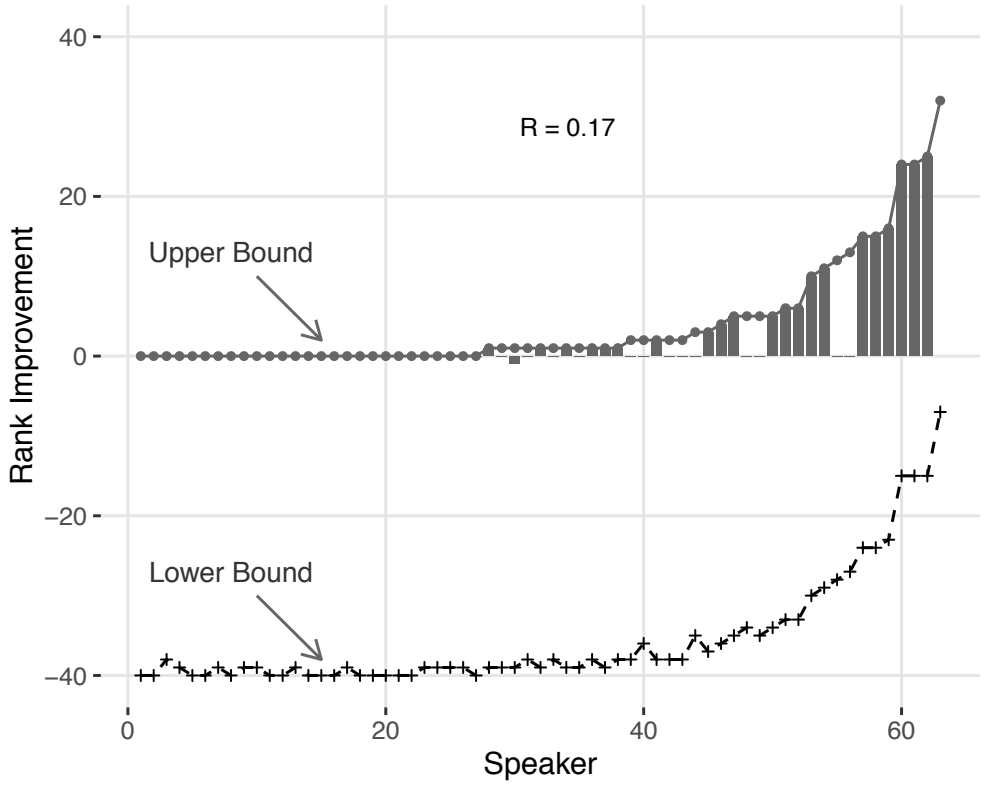

Figure 6.4: Rank improvements by using the **named variants** automatically detected in **manual** transcripts. X-axis represents the speaker instances. Y-axis shows the change in rank for each speaker instance. For each speaker instance ranked at position *i*, improvement upper bound is *i-1*, and the improvement lower bound is *number of candidates - i*, where *number of candidates* is the number of candidates for each speaker-recording instance with other known speakers removed.

Enron email collection, and then makes a list of all name variants associated with that entity. The knowledge base includes information about the frequency with which each variant was observed, recorded that in the list as well. For example, an individual could be mentioned as *John* in the Enron email collection five times, and mentioned (as a nickname) one time as *Johnny*. Therefore, the probability of that person entity being mentioned as *John* is estimated as $\mathfrak{p} = 5/6 = 0.83$. The system automatically scans the first two speaker turns in the manual transcript of each call in the test set for all name variants in the list, and then we rescore each candidate

Single Source	
Baseline	0.73
Email Social Network	0.70
Phone Social Network	0.65
Manual Channel	0.53
Estimated Channel	0.57
Name Variants	0.17
Multiple Sources	
Email Social Network & Estimated Channel	0.55
Phone Social Network & Estimated Channel	0.64
Email Social Network & Manual Channel	0.50
Phone Social Network & Manual Channel	0.56

Table 6.10: Evaluation of the Re-ranking results, evaluated by R.

as:

$$s_c^{'} = s_c * (1 + \beta \mathfrak{p}), \tag{6.4}$$

where $s_c^{'}$ and s_c are as defined above, β is a parameter that could be tuned to adjust the weight of name variant evidence (set to 1 in the experiments), and $\mathfrak{p}$ is the estimated probability (from the knowledge base) that the candidate is mentioned by that name (0.83 in the example). This works phenomenally well, substantially improving R from 0.73 to 0.17 As Figure 6.4 shows, nearly every candidate whose name was detected in the first two speaker turns out to be the true speaker. However, the use of nicknames learned from email body salutations and signatures (e.g., Johnny as a nickname for John) has only a small effect; when the system removes those nicknames from the knowledge base, R degrades only very slightly (from 0.17 with nicknames to 0.18 with only first and last names tokenized from the email headers).

There are, however, two caveats regarding this experiment. First, manual transcripts are utilized for these experiments, and the degree to which this result can

be replicated using speech recognition will depend on (1) the ability of segmenting the first two speaker turns; and (2) the ability of the speech recognition system to detect the name mentions. Task-specific tuning of the language model might help with that, since the list of name variants is available in advance.

Second, there is little ambiguity in the name variants among the set of 41 speakers (only 37 of which have associated knowledge base entries). With far larger speaker sets, effective techniques for disambiguation would become important. Results from entity linking in email [46] indicate that this is an entirely tractable problem (when social network evidence and evidence from content are used together), but of course both the social network and the content evidence is generally less accurately observable in speech than in email text. The use of nicknames learned from the email collection doesn't help much at all since in the test collection there is only one person "Stanley" referred to by a nickname "Stan". However, the usefulness of nickname matching to might be expected to increase when there are more people involved in a larger collection.

6.2.4.4 Combination of Multiple Sources

Table 6.10 also shows the effect of fusing the re-rankings with simple score summation. For these experiments, only the combinations of one type of social network with one type of channel information are explored, yielding four fusion pairs. The measure R improves when the email social network is used together with channel information, when compared to the already-good results from for channel

information, both for manual channel information (from 0.53 to 0.50) and for self-trained estimated channel information (from 0.57 to 0.55). No similar improvement is seen from using the telephone social network.

6.2.5 Discussion

In considering the collective results of all the above experiments, there are a few overall impressions. First, it is clear that the incorporation of social network evidence helps a little and that channel information helps somewhat more. Among the two social networks tried in this thesis, using evidence from communication patterns in the email network results in consistent improvements, both with and without the the complementary evidence from channel information. The telephone social network is even more helpful than the email network when used alone, but when used in combination with the channel information from either source the telephone social network yields no further improvement over using channel information alone. One plausible explanation for this is that all channel information, manual or automatic, ultimately relies on acoustic evidence, and acoustic evidence also informs the estimate of the telephone social network. When combining evidence, the email social network is thus a better choice as a complementary source of evidence.

The fact that channel information consistently outperformed social network information as a side feature is intriguing, but the structure of this test collection (with telephone lines used by specific people being recorded) is particularly well suited to the use of channel features. In other applications (e.g., cases in which

Table 6.11: Using speaker identification results to improve person linking for Enron phone recordings, measured by MRR.

Condition	Speaker	Social Network	Non-NIL	NIL	All
Manual Transcripts					
1	*None*	*None*	0.488	0.062	0.496
2	*Acoustic*	*None*	0.541	0.212	0.525
3	*Acoustic + Channel + Social network*	*None*	0.564	0.059	0.542
4	*Acoustic*	*Acoustic*	0.646	0.049	0.620
5	*Acoustic + Channel + Social Network*	*Acoustic + Channel + Social Network*	0.674	0.050	0.648
6	*Ground truth*	*Ground truth*	0.799	0.432	0.783
ASR Transcripts					
7	*None*	*None*	0.464	0.058	0.441
8	*Acoustic*	*None*	0.519	0.204	0.502
9	*Acoustic + Channel + Social Network*	*None*	0.536	0.057	0.509
10	*Acoustic*	*Acoustic*	0.601	0.045	0.569
11	*Acoustic + Channel + Social Network*	*Acoustic + Channel + Social Network*	0.630	0.047	0.597
12	*Ground truth*	*Ground truth*	0.753	0.403	0.713

trunk lines are recorded) it may be social network features that are of greater use. The results do show, however that it is possible to estimate channel assignments from acoustic evidence sufficiently reliably to be useful, and to achieve results close to what manual annotation was able to achieve.

One final observation is that both of the automatically-derived sources of information (the telephone social network and the estimated channel information) offer the promise for a double benefit from future improvements to acoustic speaker recognition techniques, since both automatically derived sources leverage acoustic speaker recognition. So not only will the acoustic baseline improve, but better estimates will be made on the unlabeled data, thus possibly resulting in better side information as well.

6.3 Using Speaker Identification to Improve Entity Linking

Table 6.11 shows the MRR of linking person mentions detected from the evaluation set of Enron phone recordings to the collection-specific person knowledge

base built from the Enron emails using **manual** and **ASR** transcripts. For most of the Enron phone recordings (1,731-57 =1,664), the speakers are unknown, thus the social network of the speakers in the collection is unavailable for the entity linking system. In the table, the column **Speaker** shows different ways of providing speakers to the entity linking system, including the baseline no speakers (*None*), predicted speakers using acoustic evidence only (*Acoustic*), predicted speakers using acoustic evidence and then improved by self-trained social and channel information (*Acoustic + Channel + Social network*), predicted speakers using acoustic evidence and then improved by self-trained channel and social network information (*Acoustic + Channel + Social Network*), and the manually recognized ground truth speakers (*Ground Truth*). The column **Social Network** shows different ways of providing the social network to the entity linking system, including the baseline no social network (*None*), social network built from predicted speakers using only acoustic evidence (*Acoustic*), social network built from predicted speakers using acoustic evidence and improved by self-trained channel and social network information (*Acoustic + Channel + Social Network*), and the ground truth social network (*Ground Truth*) built from the 57 recordings with true speakers.

Since there are only 6 NIL samples in the test set, the MRR efficiency for NIL mentions are not statistically stable. Therefore, in this Section, we only focus on the **Non-NIL** and **All** mentions. Comparing with no speakers provided to the entity linking system (Condition 1 & 7), adding the *Acoustic* predicted speakers into the system (Condition 2 & 8) improves the entity linking efficacy for both **Non-NIL** and **All** mentions. The use of side information (*Channel*) improves the efficacy

of speaker identification, and then further improves the linking results, comparing Condition 3 with Condition 2, and Condition 9 with Condition 8.

Another benefit of predicting speakers for each recording in the collection is that the social network of the speakers in the whole collection can also be predicted and used in the entity linking system. By adding in the self-trained social network, the linking improvements of Condition 4 & 10 over Condition 2 & 7 are statistically significant measured by two-tail paired t-test for **Non-NIL** and **All** person mentions. Conditions 4 & 10 show the results of using the speakers and social network predicted from the speaker identification system using only acoustic information. Further improvements are gained by using speakers and social network predicted from both the acoustic and side information (Condition 5 & 11). However, perhaps due to the small size of the evaluation set, although observed on both **manual** and **ASR** transcripts, these improvements achieved are not statistically significant.

With the recognition of person named mentions from the speech, the entity linking system could be applied to identify the particular person (potentially the speaker) that is being referred to. The disambiguation of the speaker improves the speaker identification performance. Better speaker identification also leads to better entity linking results, as shown in Section 6.3. Although there is no ambiguity problem for the existing Enron Phone Recording collection, there is potentially a virtuous cycle between the speaker identification and entity linking tasks, which worth future studying.

6.4 Chapter Summary

This chapter focused on the speech collection – Enron phone recordings. The entity linking system is applied on the transcripts of the phone recordings. The system achieves similar performance as on emails when the participants and recognized mentions are provided. The recognition of the speakers for the audio recordings is important to the entity linking task. However, on most of the recordings, the speakers are unavailable, thus significantly decrease the efficacy of entity linking. A speaker identification system can be used to automatically recognize the speakers. The characteristics (e.g., noise, informal language, lack of context) of conversational speech makes it hard for the speaker identification system to achieve performance similar to that on dissemination-oriented speech. However, the side information for the conversations (e.g., social network, name variants) provides new opportunities for the task.

The second focus of this chapter is using the side information extracted from the text to improve speaker identification efficacy, and thus further improves the entity linking efficacy. A new speaker identification text collection is developed. Five approaches are explored to incorporate side information to improve performance on a speaker identification task. This chapter illustrated how the Enron conversational telephone speech collection can be used for such experiments, and the experiments show that automatic predictions can be used as a basis for social network and channel analysis to improve speaker identification. Experiments show that the improved speaker identification results improve the entity linking task for Enron phone record-

ings by providing more accurately identified speakers and a larger and more accurate social network.

Chapter 7: Conclusion

This chapter concludes the thesis (Section 7.1), discusses the limitations (Section 7.2) and future work (Section 7.3), and presents insights gained from the implications (Section 7.4).

7.1 Conclusions and Findings

This thesis studied several tasks to populate a knowledge base from conversational sources. As an initial step, to answer research questions "Can general knowledge bases be used as the linking targets for the mentions of entities in conversational sources? Are collection-specific knowledge bases needed for the entity linking task?", a test collection is built by randomly selecting email messages and linking the named mentions to the general knowledge base Wikipedia. The results show that approximately two-thirds of the mentioned person entities and approximately half of the mentioned organization entities in the Enron email collection are not covered by Wikipedia. It is, therefore, potentially useful to build collection-specific knowledge bases for those entity types; location entities (for which only approximately 7% are missing from Wikipedia) seem to be less of a priority.

To answer the research question "Can collection-specific organization knowl-

edge bases be built from email collections? How well (in terms of coverage and accuracy) can collection-specific organization knowledge bases be built?", a method for automatically constructing a collection-specific organizational knowledge base from an email corpus is proposed. Information is gathered from Web sources (Google and Wikipedia) and email collection (body and signature). The results show that Google search provides the most information (68.4%) for the entities. Wikipedia and the body and signature provide information for an additional 6.7% of the organizational entities. In total, the four sources identify organizational information for 75% of the email domains.

For the entity linking task for conversational sources (emails and phone recordings), a supervised machine learning system with a large set of features is built to resolve the three types of mentions to the general knowledge base and available collection-specific knowledge bases. To answer the question "For the task of linking person named mentions to their referents in knowledge bases, what are useful sources of evidence that could be extracted from email collections, and what are effective ways of using those sources of evidence?", features are designed using both new evidence and new ways of shaping evidence. The improvement introduced by the new features are statistically significant comparing with the state-of-the-art work [31]. To answer to the research question "For the task of detecting named mentions referring to entities that are absent from all knowledge bases, what are useful sources of evidence, and what are effective ways of using those sources of evidence?" proposed in Section 1.1.2, features built for the purpose of detecting family members and detecting the absence of people who have been mentioned with full names are

found to be useful in the NIL detection task. However, the weights of those features need to be adjusted for different collections. The answers to the research question "For the task of linking named mentions of different types of entities to multiple knowledge bases, what are useful sources of evidence, and what are effective ways of using those sources of evidence?" are as follows: (1) features incorporating the conversational participants and social network information are the most useful for linking personal mentions; (2) features comparing the strings between entity name variants and the named mentions are the most useful designs for organizations and locations; and (3) to adapt the multi-KB structure, features are needed to indicate the types of entities.

As one step in building links between presently compartmentalized collaboration records, a system is proposed to link mentions of meetings found in email messages to a knowledge base of meeting entries. The collection-specific meeting knowledge base is built from the appointment entries in calendars, thus answering the research question "Can collection-specific meeting knowledge bases be built from calendars? How well (in terms of precision and recall) can collection-specific meeting knowledge bases be built?". Meeting entries referring to the same meeting are merged. The meeting linking system works well when the mentioned meeting is present in the knowledge base, although the present implementation tends to create a link when none should be made. Simulation on NIL cases shows that misleading evidence and low confidence predictions are the main issues for the failure of NIL detection, which could serve as a future research direction. Regarding the answer to the research question "For the task of linking meeting-related email messages to the

referenced meeting entries in the collection-specific meeting knowledge base, what are the useful sources of evidence, and what are the effective ways of using those sources of evidence?", the most useful features are the ones leveraging temporal information, followed by the features detecting the meeting participants and then the features based on topical similarity.

This thesis also introduced a new redistributable speaker identification test collection based on the recorded telephone calls of Enron energy traders. Experiments with these recordings demonstrate that the side information (e.g., social network features and recording channel metadata) can be used to reduce error rates in speaker identification and answer the following research question: "Can we make use of side information to improve the speaker identification efficacy for telephone speech? Can the efficacy improvement in speaker identification lead to efficacy improvement in entity linking?". Self-trained social network and channel features were found to be useful. Also the improved speaker identification leads to the improved entity linking for phone recordings.

7.2 Limitations

Regarding the construction of a collection-specific organizational knowledge base, the proposed methods were unable to provide information for one quarter of the entities, which is a limitation of the proposed knowledge base. Additional coverage might be achieved through better processing of domains, such as identifying the originators of spam. Another limitation for the collection-specific knowledge base

construction is that the proposed method is rule-based, which limits the types of attributes extracted for each entity. One possible solution is to apply slot filling task [90] introduced in 2.1 and to fill in different types of attributes automatically.

For the entity linking task, one of the limitations is that the types of named mentions are pre-defined as person, organization and location. In future work, the types of entities could be extended to other types, such as vehicles and products. An automatic named entity recognition system is applied to recognize the entity mentions for email messages. However, the named entity recognition task is particularly challenging in conversational speech due to poor ASR results. Currently the named mentions are recognized manually from poor ASR transcripts or extracted automatically from manual transcripts, which limits the development of automatic knowledge base population for speech collection. One possible solution is to automate the detection of mentions by tailoring spoken term detection techniques. Since there are typically multiple entity mentions in a conversation and since the referents of those mentions might be related, one future direction to resolve all mentions in the same conversation collectively. The named entity recognition used in this work was not designed to exploit characteristics of conversations to make the task easier (e.g., informal language and collection-specific mentions). Leveraging such features could be important because in some ways, entity discovery in email is harder than entity discovery in news.

Model-based evaluation is used to answer the research questions for entity linking tasks. For each task, an evaluation collection is first built to guide the development of the system. However, there is a limitation of the proposed work in

the stage of building reliable evaluation collections. From all of the email messages and phone recordings used in the experiments, a large proportion of the person referents are communicatively participating; they are the relatively easy queries. The task is more challenging when the referents are not in the header or even in any of the available knowledge bases. However, these are also hard queries for human annotators due to the lack of evidence and large number of potential candidates. The annotators of the Elsayed test collection report an agreement of 64% on the hard query judgments. On the Avocado email collection, the independent assessor tends to judge these queries as unresolvable. On the phone recording collection, the annotators are not able to recognize NIL named mentions due to the lack of context. One possible solution is studying the different reasons for NIL annotations in each test set and then using only the true named mentions that are referencing entities absent from the knowledge bases as the NIL annotations. Another potential solution is to provide a more sophisticated interface for the human annotators to use when searching for evidence. Of course the assessments by the participants of the conversations would be the ideal solution.

7.3 Future Work

As the number of populated collection-specific knowledge bases increases, one next step is to automatically connect the entities from different knowledge bases using relationships. The types of relationships could be pre-defined or learned automatically. Assigning each attribute or relationship in the knowledge bases with

confidence (e.g., a number between 0 and 1 to represent the credibility of this fact) will benefit future information searching and reasoning. The confidence of a fact could be defined by the frequency of observing it from different sources. Another future direction is phasing the entity linking task into two steps: automatically detecting the proper knowledge base and then linking the recognized mention to the chosen knowledge base. This idea was initially explored in [43].

For the entity linking task, one future direction is to extend the test collection for conversational speech to randomly sampled recordings from the full collection rather than relying only on the fully transcribed recordings. Another future direction is to integrate the features designed for conversations and the features that have proven to be useful for dissemination-oriented content into one single system. In the email collection, conversations exist that start from the subscription of sport or economy news. The first email containing the news is dissemination-oriented content, and the named mentions could be linked to general knowledge bases using the features designed for the broad entity linking task. The subsequent discussion may include personal names known only to the conversational participants, which could be linked to person entities by the entity linking system introduced in this thesis. The integration of the features could be the SVM model used in this thesis or another machine learning model (e.g., random forest, naive bayes).

In scientific collaboration, various information exchange platforms (e.g., instant messaging or teleconferences) are used. In future work on meeting linking, one potential direction is to integrate other sources to enhance information archiving and organizing. Beyond the person, organization and location entity linking tasks and

the meeting linking task, a next natural step is to link mentions of project-specific artifacts (e.g., samples, reports, and experiment results) to messages describing those artifacts. Another future step could be to build a new test collection to explicitly study the NIL cases and the NIL detection problem. More details for NIL annotations should be collected from human annotators. Still another future direction for the meeting linking task is a weighted evaluation method. Consider two email messages referring to the same meeting in which one email message contains a large amount of relevant information and attachments, while the other one is short and contains less information. The evaluation metric could be designed to favor the prediction of the email message with more information.

For the speaker identification task, the experiments with name-mention features using manual transcripts yielded improvements that allowed us to study the effect of adding nicknames to the set of known name variants. In addition to the productive opportunities for future work that are identified throughout this thesis, another potential future direction is to expand the size of the test set by manually annotating the speakers on randomly selected non-transcribed recordings. A larger test set would enhance the ability to detect statistically significant differences and would also allow us to create a development test set on which we could train the model parameters that are selected arbitrarily for these experiments.

Another future direction is to experiment with the use of spoken term detection for personal names, automating a process that our experiments in 6.1 with manual transcripts have shown to have substantial potential for yielding improvements. Another limitation of the current work in 6.2 is that all five types of side information

are integrated in a relatively simple way. Thus, future work could involve a machine learning method to combine all side information to achieve greater improvement. To support the proposed entity linking task, the current speaker identification task is framed as recognizing the speakers for each conversation. There is another potential line of work associated with speaker identification, which is to retrieve all the recordings that contain a particular speaker. In the Enron phone recordings collection, there are only on average 2.4 recordings that are manually judged to contain a particular speaker. The speakers are unknown for the vast majority of the recordings in the collection. Thus, for the speaker-retrieval task, the problem remains of how to judge the retrieved list with most of the items unannotated. Evaluation methods [9, 10, 42, 47, 51, 63] could be applied to address this problem.

7.4 Implications

Information retrieval systems, including the widely used commercial search engines (e.g., Google, Bing), are perhaps one of the most popular ways for people to interact with information. However, most of those widely used systems have for a long time been aiming at retrieving documents that contain the keywords, leaving the work of information analysis to the users. What is important here is that this thesis discussed a way of using the systems to automatically organize and analyze the information from unstructured data on the semantic level rather than the word level. The word "chair" in a document is not only a string, but could also represent an entity. For each entity "chair", there could be attributes such as "type:furniture"

and "color:white" associated with it in the knowledge base. With the structured knowledge, the computer is able to apply the calculating and reasoning functions on the information. By interacting with the system, the users are able to obtain the answers rather than document pieces containing the answers.

This thesis is an initial attempt towards the population of knowledge base for conversational sources with a focus on the entity linking task. Collection-specific organization knowledge bases are built through a rule-based system from email collections. In this thesis, only organizations with domains appeared in the email addresses are extracted from the email collection. However, organization is a fluid concept, from big companies (e.g., Google, Microsoft) to working groups (e.g., Doug's e-Discovery lab). Cold start knowledge base population techniques, defined as populating a knowledge base from scratch without a pre-existing external knowledge base, could be applied to extract organization entities without domains appeared in email addresses. With cold start knowledge base population ([37,38,87,88,90–93]), the knowledge bases for different types of entities can be built from scratch without specific data format requirements such as the ones in this thesis (e.g., email addresses, calendars). With the built knowledge bases, massive amount of historical archives could be organized automatically for researchers to browse and study; symptoms and treatments could be automatically extracted and linked from large amount of medical records; human communications could be automatically organized from massive evidence for lawyers to review.

The constructed knowledge bases are contributions by themselves in terms of information access and archiving. More importantly, knowledge bases can be used as

the foundation for other applications (e.g., privacy protection, personal assistant) by providing the knowledge in linked graph format. With populated knowledge bases, applications could be developed by third parties by calling the owner provided application programming interface. As a conclusion, it would be interesting to see the materialization of the methods developed in thesis in real world scenarios, both to see how flexible and accurate our systems are, and to gain further insights into the potential applications that could be built from them.

Bibliography

[1] Rafik Abbes, Karen Pinel-Sauvagnat, Nathalie Hernandez, and Mohand Boughanem. IRIT at TREC knowledge base acceleration 2013: Cumulative citation recommendation task. In *The Twenty-Second Text REtrieval Conference*, 2013.

[2] Jose-Luis Ambite, Vinay K. Chaudhri, Richard Fikes, Jessica Jenkins, Sunil Mishra, Maria Muslea, Tomas Uribe, and Guizhen Yang. Design and implementation of the CALO query manager. In *Proceedings of the 18th Conference on Innovative Applications of Artificial Intelligence - Volume 2*, pages 1751–1758, 2006.

[3] Satanjeev Banerjee and Alexander I Rudnicky. Segmenting meetings into agenda items by extracting implicit supervision from human note-taking. In *Proceedings of the 12th International Conference on Intelligent User Interfaces*, pages 151–159, 2007.

[4] Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction from the web. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence*, volume 7, pages 2670–2676, 2007.

[5] Adrian Benton, Jay Deyoung, Adam Teichert, Mark Dredze, Benjamin Van Durme, Stephen Mayhew, and Max Thomas. Faster (and better) entity linking with cascades. In *NIPS Workshop on Automated Knowledge Base Construction*, 2014.

[6] Adrian Benton and Mark Dredze. Entity linking for spoken language. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pages 225–230. Association for Computational Linguistics, 2015.

[7] Indrajit Bhattacharya and Lise Getoor. A latent dirichlet model for unsupervised entity resolution. In *Proceedings of the 2006 SIAM International Conference on Data Mining*, pages 47–58. SIAM, 2006.

[8] Ludovic Bonnefoy, Vincent Bouvier, and Patrice Bellot. LSIS/LIA at TREC 2012 knowledge base acceleration. Technical report, Laboratoire Informatique Avignon (France), 2012.

[9] Chris Buckley and Ellen M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 25–32. ACM, 2004.

[10] Ben Carterette and James Allan. Incremental test collections. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 680–687. ACM, 2005.

[11] Vitor R. Carvalho and William W. Cohen. Learning to extract signature and reply lines from email. In *Proceedings of the Conference on Email and Anti-Spam*, 2004.

[12] Taylor Cassidy, Heng Ji, Lev-Arie Ratinov, Arkaitz Zubiaga, and Hongzhao Huang. Analysis and enhancement of Wikification for microblogs with context expansion. In *Proceedings of the 23th International Conference on Computational Linguistics*, pages 441–456, 2012.

[13] Angel X. Chang and Christopher D. Manning. Sutime: A library for recognizing and normalizing time expressions. In *LREC*, volume 2012, pages 3735–3740, 2012.

[14] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[15] Xiao Cheng and Dan Roth. Relational inference for wikification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1787–1796, 2013.

[16] Jack G Conrad. E-discovery revisited: A broader perspective for IR researchers. In *DESI: Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings*, pages 237–246, 2007.

[17] Silviu Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 708–716, 2007.

[18] Sébastien Cuendet, Dilek Hakkani-Tur, and Gokhan Tur. Model adaptation for sentence segmentation from speech. In *Spoken Language Technology Workshop*, pages 102–105. IEEE, 2006.

[19] Jonathon N. Cummings and Sara Kiesler. Collaborative research across disciplinary and organizational boundaries. *Social Studies of Science*, 35(5):703–722, 2005.

[20] Jonathon N. Cummings and Sara Kiesler. Coordination costs and project outcomes in multi-university collaborations. *Research Policy*, 36(10):1620–1634, 2007.

[21] Jonathon N. Cummings and Sara Kiesler. Who collaborates successfully?: prior experience reduces collaboration barriers in distributed interdisciplinary research. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, pages 437–446. ACM, 2008.

[22] Jeffrey Dalton and Laura Dietz. Bi-directional linkability from Wikipedia to documents and back again: UMass at TREC 2012 knowledge base acceleration track. Technical report, Massachusetts University Amherst, 2012.

[23] Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. Zencrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st International Conference on World Wide Web*, pages 469–478. ACM, 2012.

[24] Christopher P. Diehl, Lise Getoor, and Galileo Namata. Name reference resolution in organizational email archives. In *Proceedings of the SIAM International Conference on Data Minng*, pages 70–91, 2006.

[25] Laura Dietz and Jeffrey Dalton. UMass at TREC 2013 knowledge base acceleration track: Bi-directional entity linking and time-aware evaluation. In *Text REtrieval Conference*, 2013.

[26] Mark Dredze, Nicholas Andrews, and Jay DeYoung. Twitter at the grammys: A social media corpus for entity linking and disambiguation. In *Proceedings of the Fourth International Workshop on Natural Language Processing for Social Media*, pages 20–25, 2016.

[27] Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. Entity disambiguation for knowledge base population. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 277–285. Association for Computational Linguistics, 2010.

[28] Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie Strassel. Overview of linguistic resources for the TAC KBP 2015 evaluations: Methodologies and results. In *Proceedings of TAC KBP 2015 Workshop, National Institute of Standards and Technology*, pages 16–17, 2015.

[29] Tamer Elsayed. *Identity Resolution in Email Collections*. PhD thesis, University of Maryland, College Park, 2009.

[30] Tamer Elsayed and Douglas W Oard. Modeling identity in archival collections of email: A preliminary study. In *Conference on Email and Anti-Spam*, pages 95–103, 2006.

[31] Tamer Elsayed, Douglas W Oard, and Galileo Namata. Resolving personal names in email using context expansion. In *Association for Computational Linguistics*, pages 941–949, 2008.

[32] Wei Feng and Jianyong Wang. We can learn your# hashtags: Connecting tweets to explicit topics. In *IEEE 30th International Conference on Data Engineering (ICDE)*, pages 856–867. IEEE, 2014.

[33] Raquel Fernández, Matthew Frampton, John Dowding, Anish Adukuzhiyil, Patrick Ehlen, and Stanley Peters. Identifying relevant phrases to summarize decisions in spoken meetings. In *INTERSPEECH*, pages 78–81, 2008.

[34] Raquel Fernández, Matthew Frampton, Patrick Ehlen, Matthew Purver, and Stanley Peters. Modelling and detecting decisions in multi-party dialogue. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 156–163. Association for Computational Linguistics, 2008.

[35] Samuel Fernando and Mark Stevenson. Adapting wikification to cultural heritage. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 101–106. Association for Computational Linguistics, 2012.

[36] Paolo Ferragina and Ugo Scaiella. Tagme: on-the-fly annotation of short text fragments (by Wikipedia entities). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 1625–1628. ACM, 2010.

[37] Tim Finin, Dawn Lawrie, Paul McNamee, James Mayfield, Douglas Oard, Nanyun Peng, Ning Gao, Yiu-Chang Lin, Joshi MacKin, and Tim Dowd. HLTCOE participation in TAC KBP 2015: Cold start and TEDL. In *Eighth Text Analysis Conference*. NIST, 2015.

[38] Tim Finin, Paul McNamee, Dawn Lawrie, James Mayfield, and Craig Harman. Hot stuff at cold start: HLTCOE participation at TAC 2014. In *7th Text Analysis Conference, Nov*, 2014.

[39] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics, 2005.

[40] John R. Frank, Steven J. Bauer, Max Kleiman-Weiner, Daniel A. Roberts, Nilesh Tripuraneni, Ce Zhang, Christopher Re, Ellen Voorhees, and Ian Soboroff. Evaluating stream filtering for entity profile updates for TREC 2013 (KBA track overview). Technical report, DTIC Document, 2013.

[41] Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 562–569. Association for Computational Linguistics, 2003.

[42] Ning Gao, Mossaab Bagdouri, and Douglas W Oard. Pearson rank: a head-weighted gap-sensitive score-based correlation coefficient. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 941–944. ACM, 2016.

[43] Ning Gao and Silviu Cucerzan. Entity linking to one thousand knowledge bases. In *39th European Conference on Information Retrieval*, pages 1–14, 2017.

[44] Ning Gao, Mark Dredze, and Douglas Oard. Knowledge base population for organization mentions in email. In *NAACL Workshop on Automated Knowledge Base Construction (AKBC)*, pages 24–28, 2016.

[45] Ning Gao, Mark Dredze, and Douglas Oard. Enhancing scientific collaboration through knowledge base population and linking for meetings. In *Hawaii International Conference on System Sciences (HICSS)*, 2017.

[46] Ning Gao, Mark Dredze, and Douglas W Oard. Person entity linking in email with nil detection. *Journal of the Association for Information Science and Technology*, 68(10):2412–2424, 2017.

[47] Ning Gao and Douglas Oard. A head-weighted gap-sensitive correlation coefficient. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 799–802. ACM, 2015.

[48] Ning Gao, Douglas Oard, and Mark Dredze. A test collection for email entity linking. In *NIPS Workshop on Automated Knowledge Base Construction*, 2014.

[49] Ning Gao, Douglas Oard, and Mark Dredze. Support for interactive identification of mentioned entities in conversational speech. In *International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 953–956, 2017.

[50] Ning Gao, Gregory Sell, Douglas Oard, and Mark Dredze. Leveraging side information for speaker identification with the enron conversational telephone speech collection. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017.

[51] Ning Gao, William Webber, and Douglas W Oard. Reducing reliance on relevance judgments for system comparison by using expectation-maximization. In *ECIR*, pages 1–12, 2014.

[52] Yegin Genc, Yasuaki Sakamoto, and Jeffrey Nickerson. Discovering context: classifying tweets through a semantic transform based on Wikipedia. *Foundations of Augmented Cognition. Directing the Future of Adaptive Systems*, pages 484–492, 2011.

[53] Ulrike Glavitsch and Peter Schäuble. A system for retrieving speech documents. In *SIGIR*, pages 168–176, 1992.

[54] Júlia Göbölös-Szabó, Natalia Prytkova, Marc Spaniol, and Gerhard Weikum. Cross-lingual data quality for knowledge base acceleration across Wikipedia editions. In *Proceedings of the 10th International Workshop on Quality in Databases*, pages 1–7. Purdue University, 2012.

[55] Craig S. Greenberg, Vincent M. Stanford, Alvin F. Martin, Meghana Yadagiri, George R. Doddington, John J. Godfrey, and Jaime Hernandez-Cordero. The 2012 NIST speaker recognition evaluation. In *INTERSPEECH*, pages 1971–1975, 2013.

[56] Stephen Guo, Ming-Wei Chang, and Emre Kiciman. To link or not to link? a study on end-to-end tweet entity linking. In *North American Chapter of the Association for Computational Linguistics Human Language Technologies*, pages 1020–1030, 2013.

[57] Cathal Gurrin, Hideo Joho, Frank Hopfgartner, Liting Zhou, and Rami Albatal. Overview of NTCIR-12 lifelog task. 2016.

[58] Cathal Gurrin, Alan F. Smeaton, Aiden R. Doherty, et al. Lifelogging: Personal big data. *Foundations and Trends in Information Retrieval*, 8(1):1–125, 2014.

[59] Umit Guz, Sébastien Cuendet, Dilek Hakkani-Tür, and Gokhan Tur. Co-training using prosodic and lexical information for sentence segmentation. In *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[60] Ben Hachey, Will Radford, and James R. Curran. Graph-based named entity linking with wikipedia. In *International Conference on Web Information Systems Engineering*, pages 213–226. Springer, 2011.

[61] Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R. Curran. Evaluating entity linking with Wikipedia. *Artificial Intelligence*, 194:130–150, 2013.

[62] Xianpei Han and Jun Zhao. Named entity disambiguation by leveraging Wikipedia semantic knowledge. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 215–224. ACM, 2009.

[63] Claudia Hauff, Djoerd Hiemstra, Leif Azzopardi, and Franciska De Jong. A case for automatic system evaluation. In *European Conference on Information Retrieval*, pages 153–165. Springer, 2010.

[64] Alexander G. Hauptmann, Jiang Gao, Rong Yan, Yanjun Qi, Jie Yang, and Howard D. Wactlar. Automated analysis of nursing home observations. *IEEE Pervasive Computing*, 3(2):15–21, 2004.

[65] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, Edwin Lewis-Kelham, Gerard De Melo, and Gerhard Weikum. YAGO2: Exploring and querying world knowledge in time, space, context, and many languages. In *Proceedings of the 20th International Conference Companion on World Wide Web*, pages 229–232. ACM, 2011.

[66] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, 194:28–61, 2013.

[67] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 782–792. Association for Computational Linguistics, 2011.

[68] Pei-Yun Hsueh and Johanna D. Moore. Automatic topic segmentation and labeling in multiparty dialogue. In *Spoken Language Technology Workshop*, pages 98–101. IEEE, 2006.

[69] Pei-Yun Hsueh and Johanna D. Moore. What decisions have you made?: Automatic decision detection in meeting conversations. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 25–32, 2007.

[70] Hongzhao Huang, Yunbo Cao, Xiaojiang Huang, Heng Ji, and Chin-Yew Lin. Collective tweet wikification based on semi-supervised graph regularization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 380–390, 2014.

[71] Heng Ji, Ralph Grishman, and Hoa Dang. Overview of the TAC 2011 knowledge base population track. In *Text Analysis Conference*, 2011.

[72] Heng Ji, Joel Nothman, Ben Hachey, and Radu Florian. Overview of TAC-KBP 2015 tri-lingual entity discovery and linking. In *Proceedings of the Eighth Text Analysis Conference (TAC2015)*, 2015.

[73] Saurabh S. Kataria, Krishnan S. Kumar, Rajeev R. Rastogi, Prithviraj Sen, and Srinivasan H. Sengamedu. Entity disambiguation with hierarchical topic models. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1037–1045. ACM, 2011.

[74] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[75] Bryan Klimt and Yiming Yang. The Enron corpus: A new dataset for email classification research. In *Machine Learning: ECML 2004*, pages 217–226. Springer, 2004.

[76] Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. Collective annotation of Wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 457–466. ACM, 2009.

[77] Anthony Larcher, Kong Aik Lee, Bin Ma, and Haizhou Li. RSR2015: Database for text-dependent speaker verification using multiple pass-phrases. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[78] Omer Levy, Ido Dagan, and Jacob Goldberger. Focused entailment graphs for open ie propositions. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 87–97, 2014.

[79] Chenliang Li and Aixin Sun. Fine-grained location extraction from tweets with temporal awareness. In *Proceedings of the 37th International ACM SIGIR Conference on Researchand Development in Information Retrieval*, pages 43–52. ACM, 2014.

[80] Chenliang Li, Aixin Sun, and Anwitaman Datta. A generalized method for word sense disambiguation based on Wikipedia. In *European Conference on Information Retrieval*, pages 653–664. Springer, 2011.

[81] Chenliang Li, Aixin Sun, and Anwitaman Datta. TSDW: Two-stage word sense disambiguation using Wikipedia. *Journal of the Association for Information Science and Technology*, 64(6):1203–1223, 2013.

[82] Yang Li, Chi Wang, Fangqiu Han, Jiawei Han, Dan Roth, and Xifeng Yan. Mining evidences for named entity disambiguation. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1070–1078. ACM, 2013.

[83] Xiaohua Liu, Yitong Li, Haocheng Wu, Ming Zhou, Furu Wei, and Yi Lu. Entity linking for Tweets. In *Association for Computational Linguistics*, pages 1304–1311, 2013.

[84] Xitong Liu, Jerry Darko, and Hui Fang. A related entity based approach for knowledge base acceleration. In *Text REtrieval Conference*, 2013.

[85] Xitong Liu and Hui Fang. Leveraging related entities for knowledge base acceleration. In *Proceedings of the 4th International Workshop on Web-scale Knowledge Representation Retrieval and Reasoning*, pages 1–4. ACM, 2013.

[86] Sameer Maskey and Julia Hirschberg. Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[87] James Mayfield, Paul McNamee, Craig Harman, Tim Finin, and Dawn Lawrie. KELVIN: Extracting knowledge from large text collections. In *2014 AAAI Fall Symposium Series*, 2014.

[88] Paul McNamee and Hoa Trang Dang. Overview of the TAC 2009 knowledge base population track. In *Text Analysis Conference (TAC)*, volume 17, pages 111–113, 2009.

[89] Paul McNamee, Mark Dredze, Adam Gerber, Nikesh Garera, Tim Finin, James Mayfield, Christine Piatko, Delip Rao, David Yarowsky, and Markus Dreyer. HLTCOE approaches to knowledge base population at TAC 2009. In *Text Analysis Conference*, 2009.

[90] Paul McNamee, Time Finin, Dawn Lawrie, and James Mayfield. KELVIN 2.0: HLTCOE progress in cold start knowledge base population, 2013.

[91] Paul McNamee, James Mayfield, Tim Finin, and Dawn Lawrie. HLTCOE participation at TAC 2013. In *Text Analysis Conference*, 2013.

[92] Paul McNamee, James Mayfield, Tim Finin, Tim Oates, Dawn Lawrie, Tan Xu, and Douglas W. Oard. KELVIN: a tool for automated knowledge base construction. In *HLT-NAACL*, pages 32–35, 2013.

[93] Paul McNamee, Veselin Stoyanov, James Mayfield, Tim Finin, Tim Oates, Tan Xu, Douglas Oard, and Dawn Lawrie. HLTCOE participation at TAC 2012: Entity linking and cold start knowledge base construction. In *Text Analysis Conference*, 2012.

[94] Olena Medelyan, David Milne, Catherine Legg, and Ian H. Witten. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67(9):716–754, 2009.

[95] Matthias R. Mehl, Simine Vazire, Nairán Ramírez-Esparza, Richard B. Slatcher, and James W. Pennebaker. Are women really more talkative than men? *Science*, 317(5834):82–82, 2007.

[96] Edgar Meij, Wouter Weerkamp, and Maarten De Rijke. Adding semantics to microblog posts. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, pages 563–572. ACM, 2012.

[97] Matthew Michelson and Sofus A. Macskassy. Discovering users' topics of interest on Twitter: a first look. In *Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data*, pages 73–80, 2010.

[98] Rada Mihalcea and Andras Csomai. Wikify!: Linking documents to encyclopedic knowledge. In *ACM International Conference on Information and Knowledge Management*, pages 233–242, 2007.

[99] Einat Minkov, William W. Cohen, and Andrew Y. Ng. Contextual search and name disambiguation in email using graphs. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 27–34. ACM, 2006.

[100] Teruko Mitamura, Zhengzhong Liu, and Eduard Hovy. Overview of TAC KBP 2015 event nugget track. In *Text Analysis Conference*, 2015.

[101] William Morgan, Pi-Chuan Chang, Surabhi Gupta, and Jason M Brenier. Automatically detecting action items in audio meeting recordings. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 96–103. Association for Computational Linguistics, 2009.

[102] Gabriel Murray, Steve Renals, and Jean Carletta. Extractive summarization of meeting recordings. 2005.

[103] Elisabeth Niemann and Iryna Gurevych. The people's web meets linguistic knowledge: automatic sense alignment of Wikipedia and Wordnet. In *Proceedings of the Ninth International Conference on Computational Semantics*, pages 205–214. Association for Computational Linguistics, 2011.

[104] Douglas Oard, Abhijeet Sangwan, and John H.L. Hansen. Reconstruction of apollo mission control center activity. In *the Proceedings of the First Workshop on the Exploration, Navigation and Retrieval of Information in Cultural Heritage, ENRICH*, 2013.

[105] Douglas Oard, William Webber, David Kirsch, and Sergey Golitsynskiy. Avocado research email collection. *LDC2015T03. DVD. Philadelphia: Linguistic Data Consortium*, 2015.

[106] Gary M. Olson and Judith S. Olson. Distance matters. *Human-Computer Interaction*, 15(2):139–178, 2000.

[107] J. Scott Olsson and Douglas W. Oard. Combining LVCSR and vocabulary-independent ranked utterance retrieval for robust speech search. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 91–98. ACM, 2009.

[108] Andrew Peck. Search, forward; will manual document review and keyword searches be replaced by computer-assisted coding. *Law Technology News (Online) Oct*, 2011.

[109] Bianca Pereira. Entity linking with multiple knowledge bases: An ontology modularization approach. In *The Semantic Web–ISWC 2014*, pages 513–520. Springer, 2014.

[110] Daniele Pighin, Marco Cornolti, Enrique Alfonseca, and Katja Filippova. Modelling events through memory-based, open-IE patterns for abstractive summarization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 892–901, 2014.

[111] Matthew Purver, John Dowding, John Niekrasz, Patrick Ehlen, Sharareh Noorbaloochi, and Stanley Peters. Detecting and summarizing action items in multi-party dialogue. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 18–25, 2007.

[112] Matthew Purver, Thomas L. Griffiths, Konrad P. Körding, and Joshua B. Tenenbaum. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 17–24. Association for Computational Linguistics, 2006.

[113] Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1375–1384. Association for Computational Linguistics, 2011.

[114] Steve Renals, Thomas Hain, and Hervé Bourlard. Recognition and understanding of meetings the AMI and AMIDA projects. In *IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 238–247. IEEE, 2007.

[115] Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tur. A keyphrase based approach to interactive meeting summarization. In *Spoken Language Technology Workshop*, pages 153–156. IEEE, 2008.

[116] Korbinian Riedhammer, Dan Gillick, Benoit Favre, and Dilek Hakkani-Tür. Packing the meeting summarization knapsack. In *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[117] Maria Ruiz-Casado, Enrique Alfonseca, and Pablo Castells. Automatic assignment of Wikipedia encyclopedic entries to Wordnet synsets. In *Advances in Web Intelligence*, pages 380–386. Springer, 2005.

[118] Abhijeet Sangwan, Ali Ziaei, and John H. L. Hansen. ProfLifeLog: Environmental Analysis and Keyword Recognition for Naturalistic Daily Audio Streams. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012.

[119] Gregory Sell and Daniel Garcia-Romero. Speaker Diarization with PLDA I-Vector Scoring and Unsupervised Calibration. In *Proceedings of the IEEE Spoken Language Technology Workshop*, 2014.

[120] Wei Shen, Jianyong Wang, Ping Luo, and Min Wang. Linking named entities in tweets with knowledge base via user interest modeling. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 68–76. ACM, 2013.

[121] Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Gökhan Tür. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1):127–154, 2000.

[122] Avirup Sil, Ernest Cronin, Penghai Nie, Yinfei Yang, Ana-Maria Popescu, and Alexander Yates. Linking named entities to any database. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 116–127. Association for Computational Linguistics, 2012.

[123] Stephen Soderland, John Gilmer, Robert Bart, Oren Etzioni, and Daniel S Weld. Open information extraction to KBP relations in 3 hours. In *Text Analysis Conference*, 2013.

[124] Zhiyi Song, Ann Bies, Stephanie Strassel, Joe Ellis, Teruko Mitamura, Hoa Dong, Yukari Yamakawa, and Sue Holm. Event nugget and event coreference annotation. In *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics-Human Language Technologies (NAACL HLT 2016). 4th Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, 2016.

[125] Andreas Stolcke, Xavier Anguera, Kofi Boakye, Özgür Çetin, František Grézl, Adam Janin, Arindam Mandal, Barbara Peskin, Chuck Wooters, and Jing Zheng. Further progress in meeting recognition: The ICSI-SRI spring 2005 speech-to-text evaluation system. In *International Workshop on Machine Learning for Multimodal Interaction*, pages 463–475. Springer, 2005.

[126] Stephanie Strassel, Mark A. Przybocki, Kay Peterson, Zhiyi Song, and Kazuaki Maeda. Linguistic resources and evaluation techniques for evaluation of cross-document automatic content extraction. In *LREC*, 2008.

[127] Fabian M. Suchanek, Johannes Hoffart, Erdal Kuzey, and Edwin Lewis-Kelham. YAGO2s: Modular high-quality information extraction with an application to flight planning. In *BTW*, volume 214, pages 515–518, 2013.

[128] Yla R. Tausczik and James W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.

[129] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics, 2003.

[130] Gokhan Tur and Andreas Stolcke. Unsupervised languagemodel adaptation for meeting recognition. In *Acoustics, Speech and Signal Processing*, volume 4, pages IV–173. IEEE, 2007.

[131] Gokhan Tur, Andreas Stolcke, Lynn Voss, John Dowding, Benoît Favre, Raquel Fernández, Matthew Frampton, Michael Frandsen, Clint Frederickson, Martin Graciarena, et al. The CALO meeting speech recognition and understanding system. In *Spoken Language Technology Workshop*, pages 69–72. IEEE, 2008.

[132] Gokhan Tur, Andreas Stolcke, Lynn Voss, Stanley Peters, Dilek Hakkani-Tur, John Dowding, Benoit Favre, Raquel Fernández, Matthew Frampton, Mike Frandsen, et al. The CALO meeting assistant system. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1601–1611, 2010.

[133] L. Lynn Voss and Patrick Ehlen. The CALO meeting assistant. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 17–18. Association for Computational Linguistics, 2007.

[134] Alex Waibel, Michael Bett, Michael Finke, and Rainer Stiefelhagen. Meeting browser: Tracking and summarizing meetings. In *Proceedings of the DARPA Broadcast News Workshop*, pages 281–286, 1998.

[135] Travis Wolfe, Mark Dredze, James Mayfield, Paul McNamee, Craig Harman, Tim Finin, and Benjamin Van Durme. Interactive knowledge base population. *arXiv preprint arXiv:1506.00301*, 2015.

[136] Shasha Xie and Yang Liu. Using corpus and knowledge-based similarity measure in maximum marginal relevance for meeting summarization. In *Acoustics, Speech and Signal Processing*, pages 4985–4988. IEEE, 2008.

[137] Tan Xu and Douglas W. Oard. Exploring example-based person search in email. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1067–1068. ACM, 2012.

[138] Yi Yang and Ming-Wei Chang. S-mart: Novel tree-based structured learning algorithms applied to tweet entity linking. *arXiv preprint arXiv:1609.08075*, 2016.

[139] Wei Zhang, Jian Su, and Chew-Lim Tan. A Wikipedia-LDA model for entity linking with batch size changing instance selection. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 562–570, 2011.

[140] Yiping Zhou, Lan Nie, Omid Rouhani-Kalleh, Flavian Vasile, and Scott Gaffney. Resolving surface forms to Wikipedia topics. In *Proceedings of*

the 23rd International Conference on Computational Linguistics, pages 1335–1343. Association for Computational Linguistics, 2010.

[141] Ali Ziaei, Abhijeet Sangwan, and John H.L. Hansen. Prof-life-log: Personal interaction analysis for naturalistic audio streams. In *Acoustics, Speech and Signal Processing (ICASSP)*, pages 7770–7774. IEEE, 2013.