LAMP-TR-150 March 2008

PERSONAL NAME RESOLUTION IN EMAIL: A HEURISTIC APPROACH

Tamer Elsayed, Galileo Namata, Lise Getoor, Douglas W. Oard

Computational Linguistics and Information Processing (CLIP) Laboratory
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742-3275
{*telsayed, namatag, getoor, oard*}*@umd.edu*

Abstract

Much of the work to date on searching email has focused on personal information management. Archival access poses new challenges, including automatic association of references to unfamiliar individuals using whatever information is available about those people. This paper describes a computational approach to that task motivated by intuitions about the ways people might explore an email collection to find that information. The proposed approach makes use of context in a flexible and adaptive manner. Two techniques for context expansion are: a mixture model that combines evidence from each context to rank candidates, and cutoff model that ranks candidates based on the closest context in which any suitable evidence was found. Both models rely on mentions that could be resolved to a common identity as evidence of the resolution. Results on three relatively small collections indicate that the accuracy of our approach performs favorable compared to the best known technique and results on the full CMU Enron collection indicate that the approach presented in this paper scales well to larger email collections.

Keywords: name resolution, context expansion, email collections, combination of evidence.

This work has been supported in part by the Joint Institute for Knowledge Discovery at the University of Maryland.

Personal Name Resolution in Email: A Heuristic Approach

Tamer Elsayed,* Galileo Namata,* Lise Getoor,* Douglas W. Oard†
University of Maryland, College Park, MD 20742
{telsayed, namatag, getoor, oard}@umd.edu

March 17, 2008

Abstract

Much of the work to date on searching email has focused on personal information management. Archival access poses new challenges, including automatic association of references to unfamiliar individuals using whatever information is available about those people. This paper describes a computational approach to that task motivated by intuitions about the ways people might explore an email collection to find that information. The proposed approach makes use of context in a flexible and adaptive manner. Two techniques for context expansion are: a mixture model that combines evidence from each context to rank candidates, and cutoff model that ranks candidates based on the closest context in which any suitable evidence was found. Both models rely on mentions that could be resolved to a common identity as evidence of the resolution. Results on three relatively small collections indicate that the accuracy of our approach performs favorable compared to the best known technique and results on the full CMU Enron collection indicate that the approach presented in this paper scales well to larger email collections.

1 Introduction

The increasing prevalence of informally written texts from which a dialog structure can be reconstructed, what might be called "conversational media" (e.g., threaded discussion lists, blogs, emails, and instant messaging), creates important new opportunities for personal information management, knowledge management, and archival access applications. Realizing this potential will require that we address a number of challenges, including modeling content in ways that accommodate informal use of language, modeling identity in ways that help searchers unfamiliar with a collection to make sense of individual roles, and development of procedural controls on the ethical use of personal information. In earlier work [11], we proposed a computational model of identity for email communications in which associations between email addresses, full names, and nicknames were discovered in the Enron email collection [17]. In this paper, we extend that work by associating mentions (e.g., first name or nicknames) in unstructured text (i.e., the body of an email and/or the subject line) to modeled identities. We see at least two direct applications for this work: (1) helping searchers who are unfamiliar with the contents of an email collection better understand the context of

*Computer Science Department, University of Maryland
†College of Information Studies, University of Maryland

emails that they find, and (2) augmenting more typical social networks (based on senders and recipients) with additional links based on references found in unstructured text.

Most approaches to resolving identity can be decomposed into four sub-problems: (1) finding a reference that requires resolution, (2) identifying candidates, (3) assembling evidence, and (4) choosing among the candidates based on the evidence. For the work reported in this paper, we rely on the user to designate references requiring resolution. Candidate identification is a computational expedient that permits the evidence assembly effort to be efficiently focused; we use only simple techniques for that task. Our principal contribution is the approach we take to evidence generation, leveraging three ways of linking to emails where evidence might be found: reply chains, overlapping participants (senders and receivers), and topical similarity. Our approach to choosing among candidates relies on a heuristic assignment of scores based on evidence combination, followed by selection of the most likely candidate. We evaluate the effectiveness of our approach on four collections, three of which have previously reported results for comparison, and one that is considerably larger than the others.

Large collections offer greater scope for assembling evidence, but they pose additional challenges as well. The most obvious challenge is that ambiguity naturally increases as the number of people in a collection grows. With more than 100,000 unique identities in the Enron collection, common nicknames might refer to any one of several hundred people. A perhaps less obvious challenge is that the evidence found in large collections is typically sharply skewed in favor of a few candidates who appear more frequently in the collection. For example, the Enron collection was assembled from the stored email folders of 150 Enron employees, at least one of them will therefore show up as a sender or recipient on almost every email. Naive approaches to entity resolution might tend to resolve references to whichever of those 150 people have the most similar name simply. Such a simple heuristic actually works reasonably well on very small collections, but it fails dismally on large collections in which the range of possible referents is quite diverse. We tried two approaches to evidence combination that were designed in ways that were intended to make use of several progressively more inclusive (but progressively less informative) contexts. Our "mixture model" takes advantage of all observed evidence, while our "cutoff model" guards against over-expansion by applying an early cutoff once informative evidence has been discovered in the tightest available context.

In this paper, we present a mention resolution approach that tries to automatically perform this kind of intuitively motivated search strategy. The approach explores multiple context spaces and automatically uses a set of recognized mentions in other related emails as a basis for ranking the set of candidate identities, one of which will hopefully be the true referent. The remainder of this paper is organized as follows. Section 2 surveys prior research that is related to our problem. Section 3 discusses the approach we adopt to model identity. Section 4 motivates our resolution approach. Sections 5, 6, 7, 8, and 9 then introduce our approach in increasing detail. Section 10 presents experiment results that show that our approach outperforms previously published techniques on relatively small collections, and that it scales well to the full Enron collection. Section 11 concludes the paper with a description of our plans for future work.

2 Related Work

Our approach to personal name resolution builds on related work on name recognition, identity modeling, and entity resolution.

2.1 Name Recognition

Named entity recognition is a well-studied problem in formal texts that are written to be accessible to a relatively broad audience and that generally follow standard grammatical conventions. Previous work has been done on news articles [14, 16, 6, 20], Web pages [12] and military reports [15, 25]. More recently, there has also been growing interest on named entity recognition in informal texts [2, 8]. For example, Maynard et al. [19] created the MUSE system as a multipurpose NER system on both formal and informal texts, including emails, and Minkov et al. [22] used structural features, name repetition and a name dictionary to identify personal names in email.

2.2 Modeling Identity

The growing availability of historically significant email collections has inspired work on helping users make sense of the identities, roles and relationships of individuals who participated in archived email exchanges. Carvalho and Cohen [7] used machine learning methods to detect "signatures" in email messages that could serve as concise descriptions of the sender, and Culotta et al. [9] searched the Web to find additional information on individuals whose name and email addresses were found in email headers. Work has also been done to identify multiple email addresses for the same person [13] and to draw multiple sources of evidence together to form more comprehensive identity models [11].

2.3 Entity Resolution

Although much has been done on entity resolution [23, 24, 5], resolution of personal names in email collections has received far less attention than better researched problems such as database merging. Abadi [1] used email orders from an online retailer to resolve name references in product orders. Their resolution, however, focused on resolving products, rather than individuals. Holzer et al. [13] used the Web to acquire information about individuals identified by names and email addresses from headers of an email collection. Our work is focused on resolving personal name references in the full email including the message body; a problem first explored by Diehl et al. [10] using header-based traffic analysis techniques. Minkov et al. [21] studied the same problem using a lazy graph walk based on both headers and content. Those two recent studies reported results on different test collections, however, making direct comparisons difficult. We have therefore adopted their test collections in order to establish a common point of reference.

3 Identity Modeling in Email Collections

3.1 Simple Model Representation

In a collection of emails, individuals often use different email addresses, multiple forms of their proper names, and different nicknames. In order to track references to a person over a large corpus, we need to capture as many as possible of these different types of attributes in one representation. Here we present the basic concepts of our model representation.

1. We model an identity as a set of attributes, much in the same way that WordNet models meaning (i.e. a word sense) as a set of words that express that meaning.

2. In the current model, we focus exclusively on referential rather than behavioral attributes because (1) that is where our intuition was strongest, and (2) it is the basis of our mention resolution technique.
3. 4 types of referential attributes are extracted: email addresses, names, nicknames, and usernames. We distinguish between names, nicknames, and usernames as follows:
   - Any string that co-occurs with the email address in the headers is primarily called a "*name*". Many forms of names are observed in the headers but the most frequent are "First Last", "Last, First", and "First MI Last".
   - We call any string that is used to refer to an identity in email free text (specifically in the salutation and signature lines) a "*nickname*".
   - "*Username*" is just a tokenized version of the substring that precedes "@" in the email address; for example a username extracted from "susan_scott@enron.com" is "susan scott". The user name is sometimes useful in absence of any other type of names.
4. Our present relatively simple model of identity is built from pairwise co-occurrence of referential attributes, i.e., co-occurrence associations. For example, an "address-nickname" association $< e, n >$ is inferred whenever a nickname n is usually observed in signature lines of emails sent from email address e.
5. Associations are weighted in a way that reflects the strength of the available evidence of observing each.
6. Since an email address is bound to one personal identity (except in rare cases), we leverage email addresses as the basis on which the seed models are built. This is achieved by mandating that at least one email address must appear in any observed association.
7. We have only four types of associations; each associates an email address with one of the four different attributes listed above.
8. We can think of the set of models as an undirected graph in which the nodes represent attributes and the edges represent associations. Consequently, one identity model is a connected component in this weighted attribute graph.
9. Each identity should have a unique identifier assigned automatically by the system (e.g., a sequence number). This identifier can be used to link behavioral attributes or inferred information that are associated more to the identity as a whole than to a specific attribute.

Figure 1 depicts an example of identity model. For this simple example, two attributes were linked whenever a co-occurrence association with "robert.bruce@enron.com" is observed. Sources and frequency of evidence are indicated on each association. Notice that for each type of association, there are specific sources of evidence. Notice that the selection of email addresses to be the corner stone in building the models facilitates subsequent merging of model instances by reducing it to simply linking different email addresses.

3.2 Computational Model for Mention Resolution

In the previous section, we designed a lower-level representation of a simple identity model, on which more sophisticated models can be built. For the purpose of resolving name mentions, which is our main usage of that model, it is crucial to compute an estimate of the system's prior belief (in general) that an identity is referred to by a given mention. This estimate plays

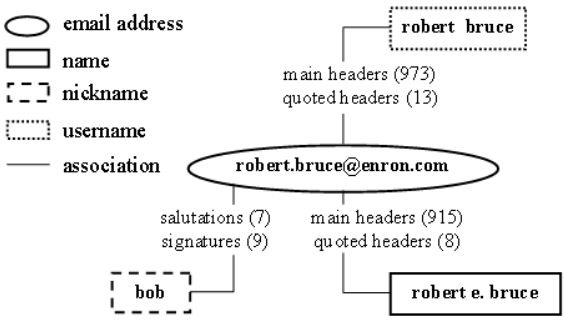

Figure 1: An example of identity model.

(along with other factors, as we explain in section 7) an important role in deciding whether a given mention can be resolved by that identity.

Intuitively, the strength of the inferred association between a name and an identity can be estimated by observing how often that association occurs in the collection, specifically in the headers or in the salutation and signature lines in the subset of emails that the person sent or received. In order to measure that strength, we postprocessed all possible name strings that appear in the model as follows:

1. We define 4 different single-token name types: first, middle, last, and nickname.
2. We normalize the different forms of observed full names such as "First Last", "Last, First", and "Last, First Middle" (sometimes surrounded by single or double quotes) to the first form with optional middle and without any quotation.
3. Each single token in that string is then labeled based on its relative position as being the first, middle, or last name.
4. We treat usernames similar to full names if it has more than one token, otherwise it is ignored.
5. The same string may appear as both a first name and as a nickname.

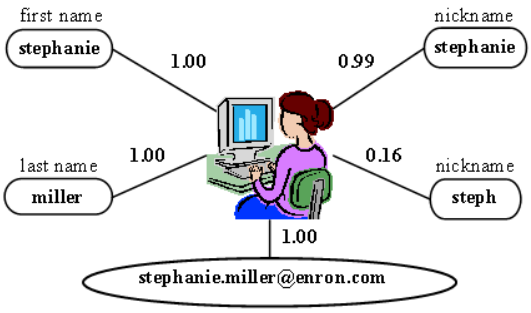

Figure 2: Higher level example of identity model.

An illustrative example of a resulting identity model is shown in Figure 2. Notice that the numbers shown on the links indicate the estimated association strength. Here we introduce a heuristically motivated estimation of the name strength that was used in our preliminary experiments of name mention resolution:

1. We estimate the association strength between a single-token name n of type t and an identity i by the fraction of the observed associations of type t that are assigned to n in the model i. If n is of type "nickname", for example, the strength is estimated as follows:

$$AssocStrength(n|i, nickname) = \frac{freq(n|i, nickname)}{\sum_{n' \in i} freq(n'|i, nickname)}$$

where $freq(n|i, nickname)$ is the total absolute frequency of n as being a nickname in all observed co-occurrence associations of i.

2. If n is a multi-token name (e.g., a full name composed of first, middle and last names), each token n_j is assumed to have a label relative to its position in the name and the overall strength (for 3-token name) is combined as follows:

$$\begin{aligned} AssocStrength(n, i) = & AssocStrength(n_1|i, first) \\ & * AssocStrength(n_2|i, middle) \\ & * AssocStrength(n_3|i, last) \end{aligned}$$

3. We can then define the general association strength of a single-token name n to an identity i, unconditioned on type, by the maximum strength over all types:

$$AssocStrength(n, i) = \max_t(AssocStrength(n|i, t))$$

4. In resolving mentions, we can leverage a name list to guess nicknames from first names. If we know that a specific first name usually has a common nickname, but this nickname was not observed in the corpus, the association strength of that nickname is estimated by the association strength of the name multiplied by a damping factor. The value of this factor for the experiments reported in this paper was arbitrarily chosen as 0.75.

5. Since one can interchange the use of first and nicknames in both headers and free text, we consider the case in which the same name token n is labeled with more than one type (e.g., "first" and "nick" for "Stephanie" in Figure 4) by modifying the above computation as follows:

$$AssocStrength(n|i, nickname) = \frac{freq(n|i, nickname) + freq(n|i, first)}{(\sum_{n' \in i} freq(n'|i, nickname)) + freq(n|i, first)}$$

This is applied similarly to first names as well. This modification will boost the strength of a nickname that is also observed as a first name, without being affected by the observance of the actual first name.

The way we compute association strength ensures that it lies between 0 and 1, but we should note that our heuristic formula for first and nicknames would make interpretation of this value as a probability inappropriate.

Our identity modeling framework allows for association of multiple email addresses. In our earlier work, we relied on address co-occurrence characteristics that are unique to the CMU version of the Enron collection. For the approach reported in this paper, we have adopted a more general technique, merging two identities whenever they share a common full name, each with a relatively strong association strength (0.8 in our experiments) to that name. Only one merging pass is performed. Since each identity model must have at least one email address, the new (merged) identity inherits the union set of these addresses. This strategy may, of course, incorrectly merge different people with common names (e.g., John Smith), but within

a single organization people will often adopt different name variants to prevent just that sort of confusion.

Our goal in this work is to resolve personal identities, but the collections which we are working with contain email addresses for mailing lists as well. As a simple expedient to guard against inappropriate attempts to build a personal identity model for a mailing list, we remove all identity models that have more that 10 different associated full names.

4 Motivating Example

Imagine that a user is searching an email collection of an enterprise to figure out how a specific decision was made. During the search, the user finds that the decision was made during the conference call referred to in the email shown in Figure 3.

Date: Wed Dec 20 08:57:00 EST 2000
From: Kay Mann<kay.mann@enron.com>
To: Suzanne Adams <suzanne.adams@enron.com>
Subject: Re: GE CONFERENCE CALL HAS BEEN RESCHEDULED

Did Sheila want Scott to participate? Looks like the call will be too late for him.

Figure 3: An email example from Enron collection.

The email clearly shows that the individuals "Sheila" and "Scott" are involved in the decision and so the user decides to look for some information about their identities. Examining the headers of the email, the mentions could not be resolved since neither "Sheila" nor "Scott" is the sender or recipient of the email. Moreover, the text of the email does not uniquely identify either person. The user extends the search to discover that this email was part of a longer discussion between a larger set of participants. Given the context of the name mentions, the user tries to find a clue in one of the emails in that discussion. If nothing is found there, the user may use the knowledge that for the email to have made sense to the recipient(s), the mention must have been unambiguous to both the sender and the receiver(s) of the email at the time it was sent. Consequently, the user may search the recent emails sent or received by those participants, trying to reconstruct the required context that may lead the user to the right persons. The user may also need to search for other emails or discussions that talk about the same specific topic around the email in question. In the email above, for example, the user may look at other emails about a GE conference call. Finally, once the user finds some evidence that provides a potential mapping of both mentions to identities, the user has the choice to stop and immediately draw a conclusion based on what found or continue to collect more evidence to gain more confidence in the mapping.

The intuitive evidence-based search strategy adopted in our example sheds the light on 4 important questions. Answers to these questions are crucial in designing a resolution algorithm that is motivated by that strategy. These questions are:

1. What kind of evidence to look for (i.e, which evidence can support the resolution of a given mention?).
2. Where to look for evidence (i.e, which emails should be searched for evidence?).
3. How to estimate the value of evidence (i.e., are all evidence of same quality or importance?).

4. How to combine evidence to rank candidates (i.e., how can scores be assigned to candidates?).

These questions outline a framework for a family of mention resolution algorithms that are evidence-based. Each different set of answers to the above questions may represent a specific strategy. In the following sections we present the answers that our proposed approach provide.

5 What evidence to look for: Additional References

Our general notion of "evidence" is any information that tends to reduce our uncertainty of the true referent. Our approach takes as evidence any recognized mention, in the body or headers of the same or related emails in context, that is lexically-different but shares at least one common candidate identity with the queried mention. The intuition is that the true referent may have also been mentioned less ambiguously in the same or related emails. For example, consider the case when the queried mention is "Sue" and there are 10 candidates in the collection whose first or nickname is "Sue". If we later observed "Sooz" (in a related email), and only two of the 10 candidates are called by that nickname, then this can be supportive evidence for one of them to be the true referent. The recognized mention in this example had common (but smaller set of) candidates with the queried mention.

6 Where to look: Contextual Space

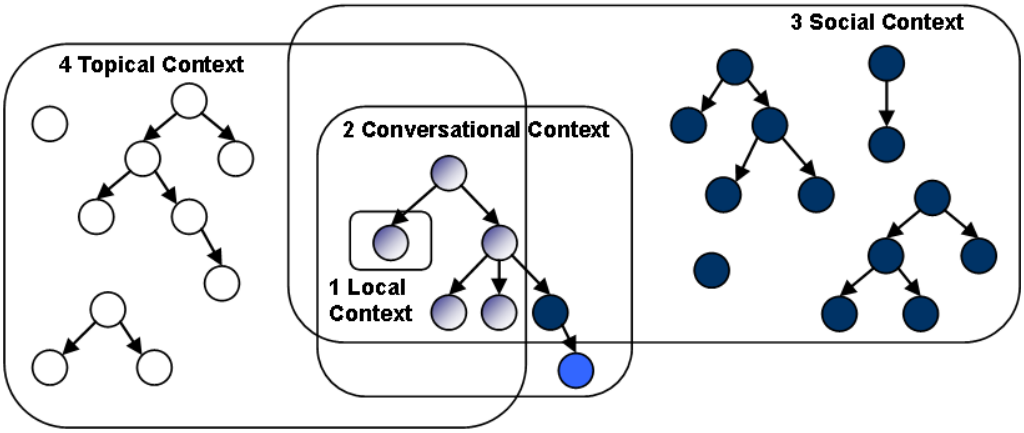

Figure 4: Contextual Space

It is obvious that understanding the context of any ambiguous mention will greatly help resolving it with high confidence. Fortunately, the nature of email as a conversational medium and the relationships between emails and people over time may reveal some clues that can be exploited to reconstruct the required context. As illustrated in Figure 4, we define 4 contexts that together comprise the contextual space of the queried mention:

1. **Local Context**: defined as the email where the named person is mentioned. This is the closest context to the mention and the starting point for reconstruction of the other contexts. We call this email the "mention-email."

2. **Conversational Context**: defined as the emails in the broader discussion in which the mention-email is contained. This expands the email context to include emails before and after the mention that is being resolved. We call this discussion the "mention-discussion."
3. **Social Context**: defined as the discussions that some or all of the participants (sender and receivers) in the mention-email have joined or initiated at around the time of the mention-email. These discussions might bear some otherwise-undetected relationship to the mention-email.
4. **Topical Context**: defined as the discussions that are topically similar to the mention-discussion that took place at around the time of the mention-email, regardless of whether the discussions share any common participants.

These generally represent a growing (although not strictly nested) contextual space around the queried mention in which evidence that can support one candidate identity over another might be found. Notice that these contexts can be further divided into sub-contexts (e.g., discussions preceding the mention-email and discussions following it within the social context).

7 How to evaluate evidence: Relevance, Specificity, and Association to Candidate

When an additional mention r is recognized, each candidate identity model which that mention could refer to is assigned credit based on a number of factors. Some of these factors are independent of the candidate such as the specific context in which r was observed, while others depend on the candidate (e.g., how confident the system is in linking r with the candidate). Here we discuss the intuition behind each factor and how we compute the overall value of observed evidence.

1. **Relevance**: One of the major factors of estimating the value of a recognized mention r is how relevant it is to the queried mention m. In particular, we should consider in which context it was observed and how far it was within that context.
   - *Context*: Evidence from different contexts should be weighted differently. For instance, a user who finds evidence in the conversational context may consider it more valuable than evidence found in the social context. Moreover, the user may prefer evidence found earlier in a direct reply chain that led to the mention-email over evidence found in different branches of the same thread. This is reflected in our algorithm by heuristically assigning different weights $W_{Context}(r|m)$ to different contexts (or sub-contexts). Table 1 shows the heuristically selected sub-context weights that we used for our experiments.
   - *Within-Context*: Within the conversational and topical contexts, we represent discussions by threads that are ranked by closeness in time, and by topic similarity, respectively, to the mention-email. Those threads should be given different weights $W_{Within-Context}(r|m)$ that are relative to their respective within-context relevance. This is applied to all but the local context.

   Consequently, the overall relevance of r with respect to m can be estimated as:

$$Relevance(r|m) = W_{Context}(r|m) * W_{Within-Context}(r|m)$$

2. **Specificity**: A recognized mention that would rule out many candidates is more useful or informative than one which would be consistent with most or all candidates in

Table 1: Weights assigned to different sub-contexts in experiments.

Local	Headers	3.00	Social	Past	1.00
	Text	2.50		Future	0.75
Conversational	Ancestors	2.00	Topical	Past	0.75
	Descendants	1.50		Future	0.75
	Others	1.25			

the set. For example, a mention of "Steph" may tend to sharpen our estimates for which "Stephanie" was intended, if "Steph" is associated with only one of the people "Stephanie" could have referred to. Our measure for this reduction in uncertainty is scaled by the number of candidates in order to produce a value between 0 and 1:

$$Specificity(r|m) = \frac{|Candidates(m)| - |Candidates(r|m)|}{|Candidates(m)|}$$

where candidates$(r|m)$ are the candidates of r that are candidates of m as well. Notice that this specificity score is zero if recognizing r would provide no new information that contributes to the resolution of m. In the future work we plan to explore information theoretic formulations of this measure.

3. **Association to Candidate**: Upon observing a mention r in some email in some context, the relative association of r to a candidate c estimates the prior belief of the system that c is referred to by r in the context of m. This estimate is computed as follows:

$$RelAssoc(r, c|m) = \frac{AssocStrength(r, c)}{\sum_{c' \in Candidates(m)} AssocStrength(r, c')}$$

Since all of the above factors intuitively contribute to the overall quality of the evidence, one way to combine them is to simply multiply their estimates:

$$EvidSupport(r, c|m) = Relevance(r|m) * Specificity(r|m) * RelAssoc(r, c|m)$$

Where $EvidSupport(r, c|m)$ denotes the credit assigned specifically to c when r is observed (i.e, the support r gave in favor of c) in the process of resolving m.

8 How to Combine Evidence: Mixture vs. Cutoff

In the situation that the system could not observe any evidence in the contextual space, the candidates should be ranked according to their prior associations to the queried mention m. This is estimated as the relative association strength of m to a candidate c over the set of candidates:

$$RelAssoc(m, c) = \frac{AssocStrength(m, c)}{\sum_{c' \in Candidates(m)} AssocStrength(m, c')}$$

If the system found any evidence then the prior belief of each candidate is scaled by the overall value of evidence that supports that candidate. We explored two ways of combining evidence from different contexts and sub-contexts. In our "*mixture model*" approach, the overall evidence support for a candidate c is simply computed as the total evidence value from *all* contexts.

$$OverallSupport_{Mixture}(c|m) = \sum_{r \in R} EvidSupport(r, c|m)$$

where R is the set of all observable evidence in the contextual space. In our alternative "*cutoff model*," we process each context in the order that they were introduced and terminate the process as soon as a threshold confidence value is reached from some subset R^* of the available evidence that has been processed up to that point. The overall evidence support for each candidate is computed in the same way as in the mixture model, but using R^* instead of R:

$$OverallSupport_{Cutoff}(c|m) = \sum_{r \in R^*} EvidSupport(r, c|m)$$

Therefore, the overall score assigned to each candidate in either model is computed as follows:

$$Score(c|m)_{model} = RelAssoc(m, c) * OverallSupport_{model}(c|m)$$

9 Implementation Details

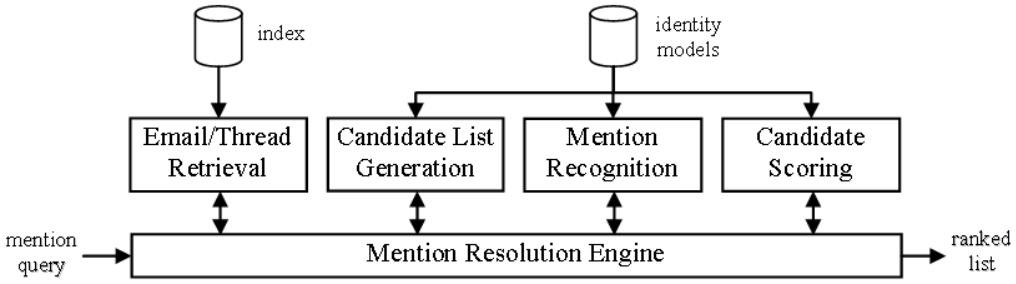

Figure 5: Mention Resolution Architecture.

Figure 5 illustrates the architecture of our resolution system. For our experiments, we assume that the user will designate a mention that requires resolution. We therefore take a specific mention as a "query" and we seek to rank candidates in decreasing order of likelihood that the mention refers to that candidate. The query consists of the mention string, a pointer to the email in which that mention appears, and the exact location of the mention in the body of that email.

9.1 Candidate List Generation

The first step in the resolution process is to generate the list of identity models that are viable candidates as the true referent. For the experiments reported in this paper, any identity model with a first name or nickname that exactly matches the mention is considered a candidate.

9.2 Recognition of Additional Mentions

For each email found in one of the four types of contexts, an efficient substring matching algorithm is used to find mentions of names. We implemented a finite state machine based on the Aho-Corasick linear-time algorithm for substring matching [3]. The tokens used to configure recognizer are all last names, nicknames, full names, and (for headers only) email addresses found in at least one candidate identity model. Sometimes, a header-extracted name includes non-name tokens (e.g., "Sue Mara at Enron SF"). To account for this, we filtered non-name tokens out using a simple heuristic based on the number of times a token appears in

the free text and headers of the corpus. Tokens that occur in more than 1000 different emails (in main body after removing salutation and signature lines) across the entire collection, but associated with fewer than 20 identity models (again, across the collection), are considered non-name tokens. Stop-words are also removed.

9.3 Context Reconstruction

The resolution engine explores four types of contexts in the search-for-evidence process. In this section, we describe how each is constructed.

1. **Local Context**: The closest context to consider is naturally the mention-email. First the address headers (From, To, Cc and Bcc) are examined to identify any reference to any candidate. The same is then done for the subject line, new (not quoted) text in the body of the email, and (if the broken link flag is set) in the quoted text within an email (headers, subject line, and body text). For each name that matches a name in a candidate identity model, the score for that candidate is incremented as described below in section 7. Recognized mentions found in free text (subject line, new body text, and possible quoted text) are counted only once per email, regardless of repetition. This same procedure is applied for every email that is examined in any context.
2. **Conversational Context**: Thread reconstruction results in a unique tree containing the mention-email. We distinguish between three subsets of emails within that thread: (1) the direct ancestors between the mention-email and the root of the tree, which capture the prior discussion context; (2) emails rooted by the mention-email, which capture the subsequent discussion context, and (3) other emails in the thread, which capture the related discussion context. We do not include the mention-email in these subsets in order to prevent double counting. Evidence from these subsets is weighted differently by our algorithm as shown in section 7.
3. **Social Context**: Discussions that share common participants but are outside a reconstructed thread may also be useful, but we expect their utility to decay somewhat with time. We therefore limit this context to threads outside the immediate discussion context that contain at least one email within a specific period of time before and after the mention-email for the experiments reported in this paper. We further distinguish between threads in which their closest email precedes the mention-email (past discussions) and those in which the closest email in time follows the mention-email (future discussions). These sets are obtained by using a preconstructed index to identify all emails that are sent or received by any of the participants in the mention-email during the specified time period. The resulting set of emails is first sorted in decreasing order of closeness in time to the time of the mention-email and then expanded to threads (removing any duplicate emails that are found lower in the time-sorted list). We remove the mention-discussion from the social context in order to prevent double counting.
4. **Topical Context**: Identifying topically similar content is a traditional topical query-by-example problem that has been well researched in, for example, the TREC routing task [18] and the Topic Detection and Tracking evaluations [4]. In our application, individual emails may be quite terse, but we can exploit the conversational structure. For the experiments reported in this paper, we tracked back to the root of the thread in which the mention-email was found and used the subject line and the body text of that root email as a Lucene query to identify topically similar emails. Terms found in the subject line are doubled in that Lucene query to give greater weight to what is sometimes a concise description of the original topic. Subsequent processing is then similar to that

used for the social context: emails from within the specified time period around the date of the mention-email are retained, ranked in this case by their topical similarity to the query, and then expanded to threads with duplicate removal, as before. We remove both the mention-discussion and discussions of the social context from the topical context in order to avoid double counting.

In both social and topical contexts, we give higher-ranked threads greater within-context weight as follows:

$$W_{Within-Context}(r|m) = \frac{1}{rank_m(thread(r))}$$

Where $thread(r)$ denotes the thread in which r was observed. For the experiments reported in this paper, we did not do fine-grained modeling of distances within the mention-email or with the mention-discussion, so this weight is set to 1 for the local and conversational contexts.

9.4 Combining Evidence

In our experiments, this cutoff model is implemented by four checkpoints, one after each type of context. If the system has found any resolution evidence when a checkpoint is reached, it treats the confidence threshold as met and ranks the candidates based on the available evidence at that point. This implementation of a cutoff model is extremely conservative, guarding against misleading evidence in less informative contexts.

For both approaches, candidates with nonzero scores are then ranked in decreasing score order. If the system finds no evidence at all, it falls back to the prior knowledge and ranks all candidates based solely on their association strength to m. In this case, every candidate will be ranked because the association strength to m is never zero.

10 Experimental Evaluation

10.1 Test Collections

We evaluate our mention resolution approach using 4 test collections, all of which are based on the CMU version of the Enron collection. Each of the test collections that we used was created by selecting a subset of that collection, selecting a set of query-mentions within emails from that subset, and creating an answer key in which each query-mention is associated with a single email address.

The first two test collections were created by Minkov et al [21]. These test collections correspond to two email accounts, "sager-e" (the "Sager" collection) and "shapiro-r" (the "Shapiro" collection). Minkov et al preprocessed the emails found in those accounts to remove ".com" emails not from Enron, emails from certain large mailing lists, forwarded and quoted portions of each email, and all addresses in the cc header field of the mention-emails. Their mention-queries and answer keys were generated automatically by identifying name mentions that correspond uniquely to individuals referenced in the cc header, and eliminating that cc entry from the header. The "Sager" test collection includes 62 mention-queries, of which we used the 51 for which Minkov reported experiment results. The "Shapiro" collection includes 60 mention-queries, of which we used the same 49 on which Minkov reported results.

The third test collection, which we call the "Enron-subset" is an extended version of the test collection created by Diehl at al [10]. Emails from all top-level folders were included in the collection, but only those that were both sent by and received by at least one email address of the form <name1>.<name2>@enron.com were retained. A set of 78 query-mentions were

Table 2: Test collections used in the experiments.

Test Collection	Emails	Threads	Identities	Queries	Avg. Candidates
Sager	1,628	1,377	627	51	4 (range 1-11)
Shapiro	974	918	855	49	8 (range 1-21)
Enron-subset	54,018	42,441	27,340	78	152 (range 1-489)
Enron-all	248,451	193,718	123,783	78	518 (range 3-1785)

manually selected and manually associated with the email address of the true referent by a graduate student in the Computer Science Department at University of Maryland using an interactive search system developed specifically to support that task. The set of query-mentions was limited to those that resolve to an address of the form <name1>.<name2>@enron.com. Names found in salutation or signature lines or that exactly match <name1> or <name2> of any of the email participants were not selected as query-mentions. Our set of 78 query-mentions include the 54 used by Diehl et al. and an additional 24 created for our experiments using the same process.

For our fourth test collection ("Enron-all"), we used the same 78 mention-queries and the answer key from the Enron-subset collection, but we used the full CMU version of the Enron collection (with duplicates removed). We use this collection to assess the scalability of our techniques. It is important to recognize, however, that all four of our collections focus the mention-queries on Enron employees, for which there is likely to be a substantial amount of evidence in Enron email.

Some descriptive statistics for each test collection are shown in Table 2.[1] The number of emails is the total number after duplicate removal and the number of threads includes trivial (single-email) threads. Different subsets yield different sets of identity models, which in turn can yield different candidate lists for the same query-mention. The Sager and Shapiro test collections are typical of personal collections, while Enron-subset and Enron-all represent organizational collections. These two types of collections differ markedly in the number of known identities and the candidate list sizes. Table 2 shows the total number of identity models after automatic construction using only information from within that collection. The table also characterizes the candidate list size, presented as an average over that collection's mention-queries and as a range. The smaller test collections naturally tend to have smaller candidate lists. Indeed, 14 of 51 mention-queries in Sager and 11 of 59 mention-queries in Shapiro have only a single candidate. Most mention-queries in the Enron-all collection, by contrast, have more than 200 candidates.

10.2 Evaluation Measures

Since each mention query has only one true referent in the answer key, this formulation of the mention resolution task is equivalent to "known item" retrieval, for which three evaluation measures are commonly used. The "*Success @ 1*" measure characterizes the accuracy of one-best selection, which can be computed from a ranked list as the mean across queries of the precision at the top rank for each query-mention. To characterize the results further down the list, we also sweep a "*Success by rank k*" measure across small integer values of k, reflecting the (averaged) availability of the correct referent to a user who is shown k alternatives. For a single-valued figure of merit that considers every list position, we use "*Mean Reciprocal Rank*" (MRR), computed as the average (across topics) of the inverse of the rank at which the correct

[1] Duplicate removal was only done for the Enron-subset and Enron-all collections.

Table 3: Summary of the accuracy results with different time periods.

	Period	MRR			Success @ 1		
	(days)	Cutoff	Mixture	Baseline	Cutoff	Mixture	Baseline
Sager	10	0.886	0.905	0.889	0.843	0.843	0.804
	100	0.889	**0.918**	0.889	0.843	**0.863**	0.804
	200	0.889	**0.918**	0.889	0.843	**0.863**	0.804
Shapiro	10	0.905	0.905	0.879	0.857	0.837	0.779
	100	0.915	0.922	0.879	**0.878**	0.857	0.779
	200	0.915	**0.935**	0.879	**0.878**	**0.878**	0.779
Enron-subset	10	0.847	0.819	-	0.795	0.744	-
	100	0.904	**0.922**	-	0.846	**0.872**	-
	200	0.910	0.921	-	0.859	**0.872**	-
Enron-all	10	**0.817**	0.766	-	**0.756**	0.654	-
	100	0.810	0.811	-	0.744	0.692	-
	200	0.812	0.813	-	0.744	0.705	-

referent is found. One way to think of MRR is as the inverse of the harmonic mean of the rank at which the correct referent is found. For example, an MRR of 0.8 would correspond to a harmonic mean of 1.25.

10.3 Results

There are five basic questions which we address in our experimental evaluation:

1. How does our approach perform compared to other approaches?
2. How is it affected by the size of the collection?
3. Which model (the mixture model or the cutoff model) performs the best, and under what circumstances?
4. How does increasing the time period (hence expanding the social and topical contexts) affect performance?
5. Which types of context make the most important contributions to the resolution task?

For this set of experiments, both cutoff and mixture models are applied with different (symmetric) time periods for selecting threads in the social and topical contexts. Three representative time periods, in days, are arbitrarily chosen: 10 (i.e., +/- 5) days, 100 (i.e., +/- 50) days, and 200 (i.e., +/- 100) days. In each case, the mention-email defines the center of this period.

A summary of the results are shown in Table 3 with the best results for each test collection highlighted in bold. Table 3 also includes (as a baseline) the results reported in Minkov et al [21] for the small collections for comparison purposes.[2] It is important to realize that the relatively small absolute differences on these smaller collections (ranging from 0.02 to 0.10) are within an order of magnitude of the differences that can be detected with this number of queries, suggesting a need to test for statistical significance before making strong claims.

[2]For "Enron-subset" collection, we have used a superset of the queries Diehl et al [10] developed, so we cannot directly compare our results to theirs. The best accuracy at the top rank (success @ 1) they reported was 0.82 for the 54 queries they used.

With these caveat in mind, it seems safe to claim that our results compare favorably with the previously reported baseline results for both Sager and Shapiro collections.

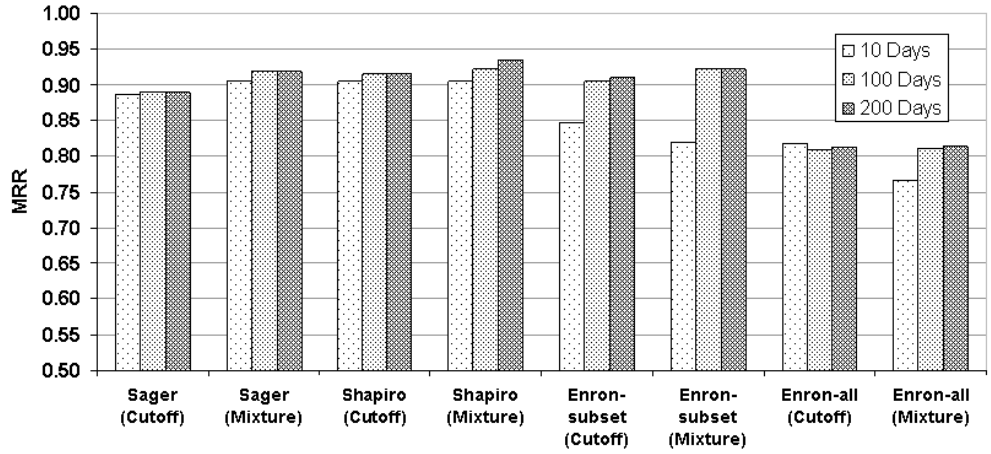

Figure 6: MRR Results for different time periods.

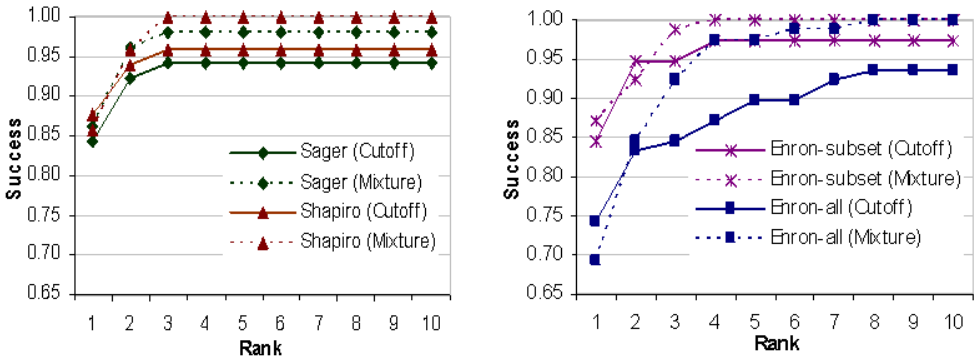

Figure 7: Success by rank k with Period set to 100 days

Our mixture model generally performed better than the cutoff model for the three smaller collections with both measures, and longer periods for time thresholds of the social and topical contexts seem beneficial in each of those cases. As Figure 6 illustrates, however, the preference for longer periods is substantial only for the two largest collections. That pattern reverses for the Enron-all collection, with the cutoff model and the shortest period being favored. Although Figure 6 illustrates the two effects seem to be coupled, only the cutoff condition benefits from the shorter periods. Moreover, Figure 7 shows the mixture model dominates the cutoff model for all four collections whenever at least three alternatives are considered. In other words, while the largest collection does indeed seem to do better with a cutoff model at rank 1 (and with MRR, which is strongly influenced by what happens at rank 1), that benefit does not extend further down a ranked list of alternatives. The observed variations with collection size seems to suggest that more sophisticated models that account for the

quantity of available evidence may ultimately be needed if we wish to design techniques that are robust across a broad range of collection sizes. We might also speculate that imperfect thread reassembly could be an increasingly important factor for larger collections, although we do not yet have evidence to support that conjecture.

Perhaps the most notable thing about our results is that they seem to be good enough for some practical applications. Specifically, our optimal one-best selection (over all conditions that we tried) is correct at least 86% of the time for the three smaller collections, and at least 75% of the time for our largest collection. Of course, the Enron-focused selection of mention-queries in every case is an important caveat on these results; we do not yet know how well our techniques will hold up with less evidence, as might be the case for mentions of people from outside Enron.

10.3.1 Contributions of Each Context

Our choice of contexts was motivated by intuition rather than experiments, so we also took this opportunity to characterize the contribution of each context to the results. We did this by using each context individually and then by using progressively more distant contexts (by our chosen order) together. We implemented this ablation study by setting some of the context weights shown in Table 2 to zero and leaving the others unchanged.

Individual Contexts Figure 8 shows the Mean Reciprocal Rank achieved with each context. In that figure, the first ("none") bar indicates how well simply choosing the identity model with the highest association strength for the query-mention string would do without looking at any evidence at all from other mentions. The difference between the two smallest and the two largest collections is immediately apparent-this ultimate fallback is remarkably effective for the smaller collections, and almost useless for the larger ones. The participant context is clearly quite useful, more so than any other single context, for every collection. This tends to support our expectation that social networks can be as informative as content networks in email collections. The topical context also seems to be useful on its own. The conversational (i.e., immediate thread) context is not very useful on its own in the larger two collections, although it is hard to guess whether this results from an infelicitous choice of weights for the three sub-contexts, from deficiencies in our automatic thread reassembly, or simply from the limited scope of the threads. The local context alone is similarly not very informative for the larger collections. It seems clear that in that case the limited scope of a single email is at least one factor in that outcome. Interestingly, in the small collections the email context proved to be fairly useful on its own despite the fact that the cc header resolving that reference had been removed.

Context Combination The principal motivation for evidence combination is that different sources may provide complementary evidence. To characterize that effect, we must look at combinations of contexts. Figure 9 shows several such combinations for the mixture model with a 100-day window (the results for 10 and 200 day periods are similar). Reassuringly, using all available evidence turns out to be a reasonable choice in every case. On the other hand, the social context alone does so well that only the Enron-all collection shows improvements in Mean Reciprocal Rank that might be noticed by a user when all available evidence is used. We should note, however, that these results were obtained with weights that we set manually before running any experiments; learned weights could yield better evidence combination.

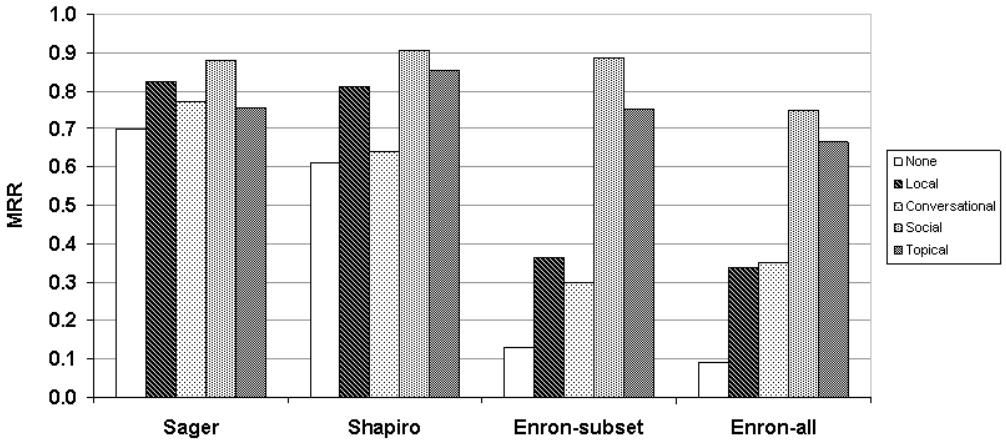

Figure 8: Mixture model with individual contexts, period set to 100 days.

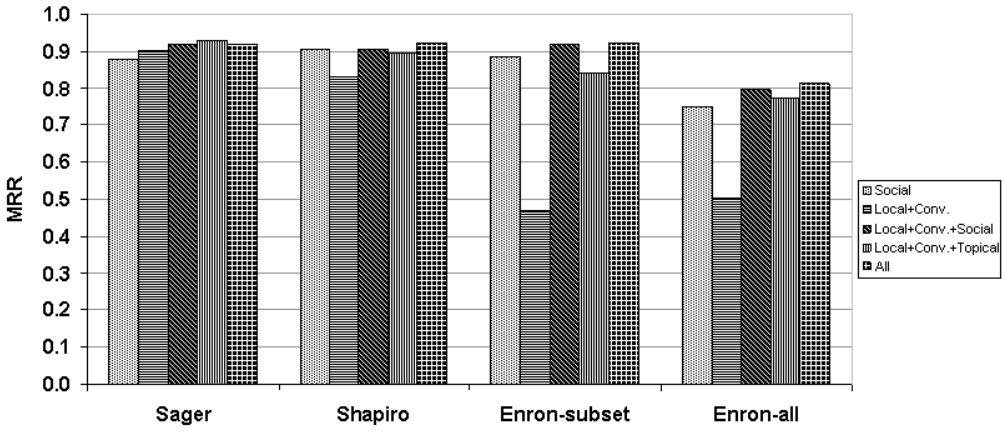

Figure 9: Mixture model with context combinations, period set to 100 days.

11 Conclusion and Future Work

We have presented an intuitive approach to name reference resolution in email which flexibly makes use of expanding reference contexts to accurately resolve the identity of a given query-mention. Our approach focuses on four naturally occurring contexts in email, including the email itself, the reply chain thread containing that email, other emails that have senders and/or recipients in common, and other emails that have significant topical content in common. Similar techniques might be applied in other scenarios where there are naturally occurring social, document and content-based contexts, such as instant messaging, multiparty chat rooms, and reconstructed reference chains between blog posts. We proposed an extended model of identity and shown how two models of evidence combination, a mixture model and a cutoff model, can use an identity model in conjunction with evidence identified in different contexts to rank potential identities for a specific name appearing in the body or subject line of an email. Our approaches outperform those previously reported in the literature, and moreover we have shown that our approach scales reasonably well to larger collections. In particular, we have seen that the potential benefit of the participant context is quite substantial in larger

collections.

In future work, we plan to extend our work to additional large email collections. Once multiple collections are available, we will gain the ability to learn parameters on one collection and then evaluate the transferability of those settings to another collection. When developing new test collections, we plan to stratify the mention-queries by evidence quantity, thus permitting us to extend our analysis to assess the accuracy of resolution strategies under progressively more challenging circumstances. For algorithm development, we are interested in exploring the potential for collective resolution strategies in which what we learn about resolving one reference can help to improve our resolution of other references in both the same and more inclusive contexts. We also plan to continue our investigation of alternative ways of modeling the name association strength and alternative evidence combination strategies. As our techniques mature, we will also need to also begin to work with real users to understand how they can best employ the technology that we create, and therefore how our technology can best be further adapted to meet their needs. We therefore see our most important contribution not as having definitively answered the questions that we asked, but rather as having posed a set questions that can motivate a rich research agenda with significant potential for improving the way people make sense of the informal interactions that may turn out to be an important part of the legacy that we leave to future generations.

Acknowledgements

This work has been supported in part by the Joint Institute for Knowledge Discovery at the University of Maryland.

References

[1] Daniel Abadi. Comparing domain-specific and non-domain-specific anaphora resolution techniques. Master's thesis, Cambridge University MPhil Masters Dissertation, 2003.

[2] Saleem Abuleil. Extracting names from arabic text for question-answering systems. In *RIAO: Recherche d'Information Assiste par Ordinateur*, 2004.

[3] Alfred V. Aho and Margaret J. Corasick. Efficient string matching: an aid to bibliographic search. In *Communications of the ACM*, 1975.

[4] James Allan, editor. Kluwer Academic Publishers, Boston, 2002.

[5] Indrajit Bhattacharya and Lise Getoor. A latent dirichlet model for unsupervised entity resolution. In *The SIAM International Conference on Data Mining (SIAM-SDM)*, Bethesda, MD, USA, 2006.

[6] Daniel M. Bikel, Richard Schwartz, and Ralph M. Weischedel. An algorithm that learns what's in a name. *Machine Learning*, 34:211–231.

[7] Vitor Carvalho and William Cohen. Learning to extract signature and reply lines from email. In *Proceedings of the 2004 Conference on Email and Anti-Spam (CEAS 04)*, August 2004.

[8] William Cohen, Einat Minkov, and A. Tomasic. Learning to understand website update requests. In *International Joint Conference on Artificial Intelligence 2005.*, 2005.

[9] Aron Culotta, Ron Bekkerman, and Andrews McCallum. Extracting social networks and contact information from email and the web. In *Proceedings of the 2004 Conference on Email and Anti-Spam (CEAS 04)*, 2004.

[10] Chris Diehl, Lise Getoor, and Galileo Namata. Name reference resolution in organizational email archives. In *Proceddings of SIAM International Conference on Data Mining*, Bethesda, MD , USA, April 20-22 2006.

[11] Tamer Elsayed and Douglas W. Oard. Modeling identity in archival collections of email: A preliminary study. In *Proceedings of the 2006 Conference on Email and Anti-Spam (CEAS 06)*, pages 95–103, Mountain View, California, July 2006.

[12] Dayne Freitag. Information extraction from html: application of a general machine learning approach. In *Proceedings of the Fifteenth Conference on Artificial Intelligence 1998.*, 1998.

[13] Ralf Holzer, Bradley Malin, and Latanya Sweeney. Email alias detection using social network analysis. In *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, pages 52–57, New York, NY, USA, 2005. ACM Press.

[14] Morris Kaufmann. *Proceedings of the Third Message Understanding Conference*. Defense Advanced Research Projects Agency, 1991.

[15] Morris Kaufmann. *Proceedings of the Third Message Understanding Conference*. Defense Advanced Research Projects Agency, 1998.

[16] Jong-Sun Kim and Martha Evans. Extracting personal names from the wall street journal. In *Proceedings of Midwest Artificial Intelligence and Cognitive Science Society '95*, 1995.

[17] Bryan Klimt and Yiming Yang. Introducing the enron corpus. In *Conference on Email and Anti-Spam*, Mountain view, CA, USA, July 30-31 2004.

[18] David Lewis. The trec-4 filtering track. In *The Fourth Text REtrieval Conference (TREC-4)*, pages 165–180, Gaithersburg, Maryland, 1997.

[19] Diana Maynard, Valentin Tablan, Cristian Ursu, Hamish Cunningham, and Y. Wilks. Named entity recognition from diverse text types. In *Recent Advances in Natural Language Processing.*, 2001.

[20] Andrew McCallum and Ben Wellner. Toward conditional models of identity uncertainty with application to proper noun coreference. In *IJCAI Workshop on Information Integration on the Web*, 2003.

[21] Einat Minkov, William W. Cohen, and Andrew Y. Ng. Contextual search and name disambiguation in email using graphs. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 27–34, New York, NY, USA, Allan20022006. ACM Press.

[22] Einat Minkov, Richard Wang, and William Cohen. Extracting personal names from emails: Applying named entity recognition to informal text. In *Human Language Technology Conference/ Conference on Empirical Methods in Natural Language Processing*, Vancouver, B.C., Canada, October 6-8 2005.

[23] Alvaro Monge and Charles Elkan. The field matching problem: algorithms and applications. In *ACM Special Interest Group on Knowledge Discovery and Data Mining 1996.*, 1996.

[24] Sunita Sarawagi and Anuradha Bhamidipaty. Interactive deduplication using active learning. In *ACM Special Interest Group on Knowledge Discovery and Data Mining 2002.*, 2002.

[25] Guodong Zhou and Jian Su. Named entity recognition using an HMM-based chunk tagger. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics 2002.*, 2002.