Bringing Structure to Personal Email Archives

Vihari Piratla
Amuse Labs
Dharwad, India
viharipiratla@gmail.com

Sudheendra Hangal
Ashoka University
Delhi NCR, India
hangal@ashoka.edu.in

Peter Chan
Stanford University Libraries
Stanford, CA 94305
pchan3@stanford.edu

Chaiyasit Manovit
Ixora Technology
Mountain View, CA
sit@ixoratech.com

Monica S. Lam
Stanford Computer Science
Stanford, CA 94305
lam@cs.stanford.edu

ABSTRACT

The personal papers of important individuals are a valuable source of material for historical research. In the digital age, much communication of historical importance is being captured in electronic archives. Recognizing this fact, special collections in libraries and museums routinely collect and preserve the personal email archives of historically important people. However, the value of these archives lies untapped because they are stored as isolated text corpora, without any connections to real-world entities or databases. We describe a system called ePADD that provides such connections, enabling interesting semantic queries over these archives.

ePADD incorporates a series of techniques to extract entities from an email archive and to resolve them to known, real-world semantic entities. We bootstrap effective classifiers tuned for a particular email corpus for various categories of entities, such as people, places, books, movies, museums, awards, universities, etc. We also look up and link these entities to well-known semantic databases such as Freebase, Wikipedia, and Library of Congress Subject Headings, ranking ambiguous matches based on how likely they are to be correct. We show that the context provided by the archive can be effective in disambiguating these entities and resolving them accurately. We have tested ePADD with email archives in our university library's special collections, and find that it is able to resolve thousands of entities in the archive, thereby enabling semantic querying of entities in such archives.

General Terms

Theory

Keywords

Information extraction, Personal data, Email, Semantic data

1. INTRODUCTION

1.1 Motivation

Letters and documents belonging to individuals are an important source of material for gaining insight into history [22]. Historians, authors, students, journalists and other researchers often refer to personal communications and papers for understanding the actions of notable individuals. Given the prevalence of electronic communications, the digital archives, especially email, of these individuals are particularly valuable for this purpose. Therefore archival organizations like libraries and museums make it a point to capture personal email archives for preservation and research.

A few points make email archives particularly valuable from a historical point of view. Like physical letters used to in the past, email captures intimate person-to-person communication. Email is used on a consistent basis by many people, and reflects all sorts of activities, whether special or mundane; it commonly contains information about people met or corresponded with, events attended, trips made, projects worked on, etc. For example, we have come across individuals who have collected well over 30 years of email at Stanford University, chronicling virtually an entire career. Just as photographs and multimedia capture rich and evocative moments from the past, the textual part of email archives complement them by capturing thoughts, emotions and feelings. In addition, emails frequently contain supplemental images and documents in the form of attachments. Another advantage is that copies of messages often exist with both the sender and the receiver, unlike paper letters where it is more difficult to get a copy of letters sent by the donor. This allows chains of conversation to be reconstructed easily. Further, email is long-lived and relatively light-weight in terms of storage; it is relatively easy and common to port email when moving from one email system to another (especially, since it is often used as a tool of record), and it is still possible to read decades-old email. Finally, due to its simplicity and open, federated architecture, email has widespread usage[1].

Email records are also often summoned in the context of litigation or Freedom of Information requests, and are of widespread interest to journalists and the wider public. The email archives of the Enron corporation, Alaska Governor

[1]Over 2 billion individuals collectively own about 3.3 billion email accounts, according to the Radicati group [18].

Sarah Palin and US Supreme court justice Elena Kagan are notable examples. The U.S. National Archives and Records Administration emphasizes that emails are federal records, and its Capstone project requires federal departments to implement automated or rules-based records management policies for email [**?**].

1.2 Applications

While it is relatively easy and increasingly common for archival organizations to collect long-term email archives, it is less easy for an archivist to process these archives and for a potential user or researcher to make use of these archives. As a result, email archives are rarely made available to researchers; instead they are often listed as a single series or sub-series in a "Finding Aid" in special collections [23]. Our work on ePADD is motivated by the difficulty we have encountered of processing large-scale archives with thousands of messages.

Consider the following hypothetical scenarios:

- A biographer is exploring the relationship between Steve Jobs and Bill Gates and wants to find people whom they might both have known.
- A historian is interested in exploring how Douglas Engelbart's work may have influenced the future of Silicon Valley, and would like to see how topics mentioned by Engelbart diffuse over time into other influential people's communications.
- A web search user is looking for information on the award-winning book Midnight's Children and would be interested to find out what is said about it in the email archives of its writer Salman Rushdie.
- A journalist wants to know which large U.S. corporations are mentioned in Alaska Governor Sarah Palin's official email correspondence.

Subject to the availability of the underlying email archives, we would like our system to enable such applications in the future. Since privacy of personal email archives is obviously a concern, some applications may be enabled only after appropriate permissions are obtained. However, the list of entities (people, places, organizations, etc.) mentioned in personal correspondence are often made public, though what exactly is said about them is restricted. For example, the traditional "finding aids" using in libraries publicly list the metadata and entities mentioned in personal letters; approved access to a library reading room is needed to see the actual letters. See Publishing entity mentions in email archives (subject to the approval of the donor or their representatives) is along the same lines.

We would also like to perform such linking in a scalable and mostly automatic fashion because email archives tend to very large, running into thousands of messages. In some cases, there has been a massive effort by newspapers and other journalistic organizations in acquiring the data and creating an interface (for example, with the Sarah Palin and Elena Kagan email archives) for people to access and annotate it [24, **?**, 17].

1.3 The ePADD system

In this paper, we describe the open-source ePADD [2] system which addresses some of the challenges above. Its key features are the following:

- It performs a brute force lookup of every possible named entity in the text to DBpedia. It differentiates between the casual and serious mention of an entity based on context overlap between email and Wikipedia page. For example in the sentence Wish you a Happy Christmas, figuring out the sense of Happy Christmas as casual mention or an entity(book,movie or poem).
- It automatically recognizes named entities in email messages using a combination of looking up seed entities generated in the above step, and training a Named Entity Recognizer. We show that this combination of techniques works much better than either technique alone.
- It links these entities (optionally under user control) with well-known semantic web databases such as Library of Congress Subject headings, Freebase and DBpedia. This allows powerful queries by search engines for the semantic web, and could conceivably be linked with structured entity graphs used by commercial search engines.

1.4 Contributions and Outline

Our major contributions in this paper are summarized below.

- We propose information extraction techniques to link personal corpora like these email archives to known real-world entities, thus allowing semantic queries to be performed on the data embedded in them. While our techniques are currently applied to email, they may be applicable to other forms of personal text such as social media and personal documents as well.
- We have built a novel, open source system called ePADD for processing email archives in special collections that enables effective processing of these email archives [3].
- We report experiences with applying these techniques to the email archives of the well-known poet Robert Creeley which are housed in Stanford University library's special collections.

To our knowledge, ePADD is the first system to link personal corpora with external entities and evaluate them on real-world email archives.

The rest of this paper is organized as follows. Section 2 provides a brief overview of the archiving process for those who may be unfamiliar with it. Section 3 surveys related work and Section 4 describes our system, techniques and

[2] ePADD is an abbreviation for Email: Processing, Appraisal, Discovery and Delivery

[3] The software system along with source code will be submitted as supplemental materials along with the archival version of this paper.

rationale for choosing them. Section 5 presents our results and experiences with using ePADD with a real-life email archive of the well-known American poet Robert Creeley. Section 6 concludes and discusses some avenues for future work.

2. RELATED WORK

The idea of a personal digital archive that is cross-linked with "associative trails" goes back to Vannevar Bush in 1945 [3]. Several projects in the archives community have already recognized the importance of email archives for historical research and are actively working on defining best processes to deal with them [1, 25, 27].

2.1 Visualization and browsing

There has been prior work on systems for mining and visualization systems for email and other text corpora [26], [13], [10]; however, they do not specifically focus on entity identification and linking. Projects like Overview and DocumentCloud are popular in the journalists' community for processing textual corpora but are not focused specifically on email archives. Newspapers have attempted to use crowdsourcing to identify interesting information in email archives, for example with the Sarah Palin emails [9, 17].

2.2 Named Entity Recognition(NER)

Basic named entity recognition techniques use machine learning to identify and classify entities into a pre-specified set of classes. Most typically, the types of entities recognized are people, places and organizations. The systems used can range from supervised to semi-supervised [20, 5] to unsupervised [4, 21, 7]. Robust open-source libraries like Open NLP and Stanford NLP implement named entity recognition techniques. A key issue that impacts performance is the data used for training; models trained with text from one domain may not do well on text from another domain. ePADD attempts to perform both training and recognition on the same corpus, thus mitigating this problem.

2.3 Fine-grained entity recognition

There has been recent research focused on fine-grained entity recognition, i.e., recognition of entities in very specific categories. Nadeau uses a semi-supervised method that starts with a few manually-entered seed instances of a type of entity, and then expands it to a more exhaustive list using relevant web documents [16]. Ling and Weld's FIGER system recognises and classifies entities into types that are compiled by cleaning and merging Freebase types [12]. Lin et al's system [11] is trained on linkable noun phrases (entities that map to Wikipedia) and can then identify unlinkable noun phrases (entities with no corresponding Wikipedia page) in the corpus. ePADD also performs fine-grained entity recognition, but uses categories that are likely to be the most relevant to researchers trying to make sense of email archives.

2.4 Disambiguation to Wikipedia

Named entity linking systems (for example, GLOW [19], Wikify [14], and others [2, 6, 15, 8]) identify and link relevant entities in free text to their corresponding sources in Wikipedia. There are many different approaches, but broadly, some context related to an entity in the corpus is computed and matched with the resolution candidate's Wikipedia page using a coherence function. The methods are classified as local, where each occurrence of an entity is resolved separately, or global, in which all occurrences of an entity are resolved jointly. ePADD employs a similar approach to resolve entities to Wikipedia, but also uses other sources.

3. ARCHIVAL PROCESS

This section provides an introduction to how email archives fit into the special collections process, and how ePADD handles each step in the process.

Table 1 shows an overview of the major phases in acquiring, processing and making usable an email archive, with the typical actions in each phase. Of course, some of the boundaries between the phases as well as who perfoms them depends on the specific situation. For example, some donors may be willing to perform some processing themselves. In other cases, a donor is deceased and a bulk of the appraisal and redaction functions have to be performed by an archivist at the special collections department. A particular donor may not allow a public discovery mode at all, and if the archive is public, the full message contents can be provided in the discovery mode as well. Of course, there are use cases for ePADD outside the special collections process as well.

While ePADD supports all phases of this table, including email collection, cleaning, de-duplication, message redaction and embargoes, visualization, browsing, search, etc., we focus in this paper on the processing and discovery phases, where the personal archive is linked to external semantic entities (often called "authority records" in library parlance). Without the help of automatic tools like ePADD, these steps are time-consuming, taking months, and are often left undone. This in turn means that potential researchers cannot make use of the valuable content in the archive without investing the time and expense of making a trip to the archive's reading room and then wading through a lot of material, a prohibitively difficult proposition. In our experience, publishing linked entities is a good balance between making the archive useful, and keeping the bulk of it confidential.

The Robert Creeley Archive: To illustrate the differences in processing email archives using the traditional approach, consider the archives of noted American poet Robert Creeley, whose archives are housed in our university's special collections, and contain both paper and digital materials. There are 7,000 letters in the paper component of this archive. In the finding aids for this archive, the correspondence listing takes 122 pages out of a total of 251 pages, indicating the importance of letters. Note that this listing had to be painstakingly and manually generated by an archivist over many months. Mr. Creeley's email corpus is even larger – consists of over 80,000 pieces of email, spanning about 13 years. The messages are loosely organized with relatively little folder structure, and with many duplicates; of these, about 49,000 messages are unique. We will use this archive to illustrate many of ePADD's features in the next two sections.

4. THE EPADD SYSTEM

Phase	Performed by	Description
Appraisal	Donor	The donor uses ePADD to gather and appraise the contents of his or her email archive and redacts sensitive messages or places embaroges restricting a message for a period of time or until some other condition is met. and transfers the archive.
Processing	Archivist	The archivist receives the archive and performs various structuring steps to identify and disambiguate entities, and to link them to authority records from sources such as DBpedia, Freebase and Library of Congress subject headings. The archivist prepares a discovery and a delivery version of the archive, along with appropriate finding aids and data dumps in linked open data format to help researchers.
Discovery	Potential researchers, public	A public version of the archive is prepared in which only the named entities, along with links to resolved authorities in external semantic databases, and message metadata like participant names and dates are available (subject to the approval of the donor), but the rest of the message contents are hidden. Potential researchers can issue semantic queries or browse the archive to discover information in the archive.
Delivery	Researchers	A researcher who has discovered something of value in the archive approaches the host organization and gains full access to the full archive, included the resolved authorities and the full text of messages, typically in a controlled reading-room environment.

Table 1: The various steps of the special collections process and how ePADD fits into them.

Given an email archive, how can we identify entities of different types that may be of particular relevance to researchers? In this section, we describe the features of ePADD that are most relevant to answer this question. ePADD breaks this down into 4 components.

- Identifying entities of various types from email text. This includes not only generic ("ENAMEX") types such as people and places, but also fine-grained types such as books, movies, universities, companies and awards that are likely to be of interest to researchers.
- Disambiguating a reference to an entity that may refer to one or more "authority records". The authority records could come from different sources such as FAST, DBpedia, Library of Congress subject headings, etc. They could also be internal authority records, such as links to entities within the address book of correspondents in the email archive itself, or to a library catalog.
- A user interface to allow an end-user such as an archivist to scan a ranked list of possible entity linkages and to select and confirm one of them,
- Export of approved entities in an RDF format to make the records available for semantic querying.

We elaborate on each of these components in the subsections below. Many of these features were implemented in response to the inadequate performance of existing tools and techniques. All the examples that follow are taken from the email archives of Robert Creeley, mentioned above.

4.1 Entity Recognition and tagging

We initially tried to identify entities using the simple approach of using the OpenNLP library with its default training models to identify people, places and organisations. These models are trained on annotated news articles. However, the performance (especially the recall) of this approach is very poor, since the language used in email has evolved in its own way – it differs from the language used in newspapers and is frequently informal, colloquial, and uses its own abbreviations and emoticons. Moreover, the language may change even from one archive to another – a poet's archive may use very different language from a sportsman's. We also need to infer many more fine-grained categories such as books, movies, museums, universities, companies, etc. that may be of potential interest to a researcher. Therefore, it is best if the recognition models are trained on the same corpus without requiring manual intervention.

To tackle this problem, ePADD first aims to identify entities that it can confidently assign to the various types of our interest. We call these high-confidence entities *seed instances*. Seed instances can be identified done with the following two step approach:

- We first identify all possible multiple word entity candidates (of all types) using the following pattern: Initial capital words connected by stop words but not ending with a stop word. This is similar to recognising a singular noun from part-of-speech tagging.
- Next, ePADD looks up each entity candidate in the DBpedia database (which also has associated categories for most entries) using case-insensitive string matching. To ensure that it picks up only high confidence matches, ePADD ignores matches with DBpedia entities that are associated with multiple categories. For example, the entity *Cricket* could match either *Cricket_(Game)* or *Cricket_(Insect)*, but in this step we match neither because we want only high confidence instances in every category for later training.

Because our candidate entity detection is fairly aggressive,

we find that the second step still generates spurious matches between things that are not really entities at all, but still match DBpedia. Consider the sentence *"He and I do apologize."*, which extracts the candidate entity *He and I*. It so happens that this is the title of an unambiguous book entry in DBpedia, and this would wrongly get marked as a seed instance. To avoid these accidental matches, we further check if some context associated with the candidate term in the email archive is also on the Wikpedia page corresponding to the DBpedia entity. Context for the candidate term is simply defined as a list of all multi-word Person, Organisation and Location entities that co-occurred with it in any email message. If there were a true match, we would expect surrounding entities to refer to the name of the author of this book, name of the characters, places associated with it, etc. In the above case, the Wikipedia page for the book *He and I* has no overlap with the entities associated with this phrase's usage in the archive, and hence the phrase is not used as a high-confidence seed instance.

The steps above generates high-confidence seed instances of each type that we are interested in. We can use these seed instances to train a model tuned for this corpus that can recognise other entities of the same type. These may be entities that are not prominent enough to have a Wikipedia page (though may exist in other databases); for example not all books or companies have a Wikipedia page. Or they may be entities that are present on Wikipedia, but are referred to with variant names like acronyms, short names, nicknames, etc. For example, in the sentence, *"She graduated from William's."*, we would like to recognise *Williams* as a university, after training, even though it would not be a direct string match in a database.

To generate the training sentences, ePADD looks up the seed instances in all email messages, and all sentence containing any seed instance are annotated with the appropriate type. In order to get good recognition performance, however, we also have to generate unlabelled data, i.e., sentences with no occurrences of the desired-type. To identify sentences with a very low probability of actually containing an entity of the desired type, we pick up all sentences in email messages that contain no type-dependent keywords. These keywords are synsets or modifiers used with the entities of the class. For example if we were looking for unlabelled sentences of the book type, we would exclude messages with any mention(s) of the words *book, volume, novel*, etc. Typically, less than 15 determinative keywords for a class can give reasonable results. We find that providing both these kinds of sentences is necessary to build a model that can identify instances of each type with good precision.

The results of running this recognition over a few types of entities are presented in the next section.

4.2 Entity disambiguation

There are 2 kinds of entity disambiguation which ePADD performs. The first kind of disambiguation relates to identifying different entities of the same type that have the same name.

4.2.1 Disambiguation to Knowledgebase

Given an entity *Charles Bernstein*, does it refer to *Charles Bernstein*, the composer, or *Charles Bernstein*, the poet? If the entity has multiple hits in the db, we intersect context between Wikipedia page and sort the possible resolutions.

A name may correspond to multiple entity records in external database. An entity record is a definite record in the database. Assigning authority to an entity involves selecting the right entity record. An archivist is required to chose an entity record in order to assign authority. There is a UI to let the user confirm authority records. A snapshot of the UI is shown in the figure 1. So as to ease the process, entity records linked to Wikipedia are resorted based on context matching between Wikipedia content and context of entity from emails.

4.2.2 Internal disambiguation

Many email messages are exchanged between people who know each other well; therefore there is a strong contextual backdrop to their communication and a kind of cryptic or short-hand notation suffices to convey meaning.

Consider a researcher who is browsing a large-scale email archive and encounters a reference to *Charles* in a message between Alice and Bob. While the reference may be immediately clear to Alice and Bob, a researcher would have to painstakingly develop a mental model of the relationship between Alice and Bob to infer which of the possibly many Charles the message might be referring to. ePADD attempts to help the user in this task by identifying and scoring likely resolutions of such single-word person names in email messages.

ePADD considers a epadd system, a multi-word person name is considered unambiguous and every mention of such entity is assumed to have only one meaning; at-least in the context of email archive. Likewise any single-word person name is considered ambiguous. We use the term "internal authority" interchangeably with multi-word person names.

We try to assign internal authority to any single-word person name to build a strongly connected internal network of personal mentions that makes traversal of email documents easy for researchers and also resolving entities such as in the figure 2 to internal authority, can be helpful for researcher for its quick tip.
Given an single-word person names in an email, multi-word candidate person names are ranked in the decreasing order of confidence when the user hovers on a term.

Candidate resolutions are the multi-word names which contain the single word name that is being resolved. For example in the figure 2, *"Peter"* is the ambiguous entity to resolve; *"Peter Gizzi"*, *"Peter Hare"*, *"Peter Novick"*, *"Peter Conners"*, *"Peter Quatermain"* are some of the candidate resolutions. For every multi-word personal name an affiliation score with every email address it appeared with and every entity it co-occurred with is pre-computed. A candidate resolution for a single word name is given a score based on co-occurring entities and correspondents in the current email. Prior probabilities of every multi-word entity is pre-

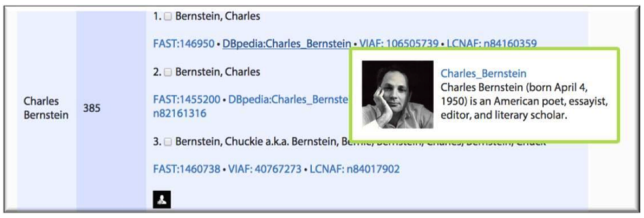

Figure 1: Each correspondent is resolved to a ranked list of possible matches in the FAST database.

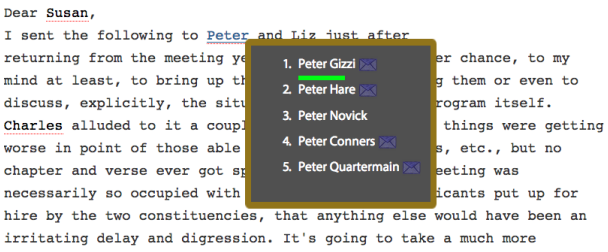

Figure 2: Hovering on a single word name in a message brings up a ranked list of suggested entity completions.

computed as a ratio of number of occurrences of the multi-word entity to total number of occurrences of any entity in all emails.

The scoring algorithm for a candidate resolution follows: Email Addresses are referred to as EA and Co-occurring entities as CE in what follows. For every ambiguous entity that is to be resolved in an email, a signature feature is constructed from correspondent addresses and unambiguous entities in the email. A candidate resolution is scored based on its affiliation with the correspondents and unambiguous entities in the current email.

let a be ambiguous entity to be resolved

Figure 3: Hovering on acronym in a message brings up a ranked list of suggested abbreviations.

Total messages	49,646
Incoming messages	30,155
Outgoing messages	19,491
Time range (year)	1994-2006
Unique Correspondents	24,010
Correspondent names	37,500

Table 2: Statistics of Robert Creeleyś Email Archive

Entity class	seed instances	recognised entities	Precision
Book	102	569	0.89
University	471	1351	0.91
Museum	60	201	0.62
Company	143	249	0.75
Award	13	34	0.55
Movie	157	325	0.59

Table 3: Results with different entities with precision

let e be candidate entity resolution(internal authority) for a let prior probabilities of entities be denoted by $entitiesPrior$ let features be string indexable data-structure which contain affiliation score of every internal authority to email addresses and co-occurring entity.

$$eScore(e|a) = \frac{\sum_{ea \in features(a).EA} features(e).EA(ea)}{\|features(e).EA\|}$$

$$cScore(e|a) = \frac{\sum_{ce \in features(a).CE} \frac{features(e).CE(ce)}{entitiesPrior(ce))}}{\|features(e).CE\|}$$

$$score(e) = (0.5 * cScore + 0.5 * eScore) * entitiesPrior(e)$$

Along with person names we also found this method to be particularly useful to also resolve acronyms. The algorithm differs from method shown above only in the candidate entity generation. The results with acronyms are shown in the figure 3

5. RESULTS

This section gives an overview of the testing data(Robert Creeley's email corpus) and discusses results and numbers in detail.

Entity class	linkable entities	non-linkable entities	Non-entities (error)
Book	116	397	44
University	405	827	92
Museum	52	74	73
Company	88	100	60
Award	11	9	9
Movie	95	97	127

Table 4: Statistics on Linkable and Non-linkable entities for different types

We evaluated our technique on the email collection of Robert Creeley, statistics of the email collection are shown in the table 2

Candidate named entities are generated as described in the section 4, 208,088 candidate entities are recognised in the entire archive; of which 8,663 entities map to a DBpedia topic(topics with disambiguation pages are not considered in this mapping for accuracy). Entities of desired type are further filtered by context matching as described in the section 4 to generate seed instances. Train samples are generated and trained to generate a model. We used maximum entropy model in our experiments.

table 3 shows statistics relating to number of seed instances and number of recognised entities along with its precision. The table 4 shows the statistics about the number of linkable(entities with corresponding wikipedia page) non-linkable(entities that has the desired type but no corresponding wikipedia page) entities. and non-entities: entities recognised by tagger but not of the desired type or is just random. Number of non-linkable entities shows the strength of the method in recognising new entities other than those initialised with.

6. LIMITATIONS AND FUTURE WORK

Currently, ePADD can smoothly handle personal archives with about 100,000 messages. Our future plans are to improve scalability and to provide cross-collection search so that library patrons can search multiple collections at once.

7. CONCLUSIONS

We have shown how long-term email archives can be processed relatively efficiently, and how they can be made partially available to the general public. Our experience with the Creeley and Fikes corpora and the resulting system should be useful to other people who need to process large-scale email archives. Our system is publicly available at the URL http://suif.stanford.edu/~hangal/epadd.

We hope tools such as ePADD will make it more common for curators to capture email archives as valuable documents of record. Currently, this process is limited by the cost of acquisition, processing and delivery.

8. REFERENCES

[1] AIMS Work Group. AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship. *White paper*, October 2011.

[2] R. C. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *EACL*, volume 6, 2006.

[3] V. Bush. As We May Think. *Atlantic Monthly*, 176, 1945.

[4] M. Collins and Y. Singer. Unsupervised models for named entity classification. In *Proceedings of the joint SIGDAT conference on empirical methods in natural language processing and very large corpora*. Citeseer, 1999.

[5] A. Cucchiarelli and P. Velardi. Unsupervised named entity recognition using syntactic and semantic contextual evidence. *Computational Linguistics*, 27(1), 2001.

[6] S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic, June 2007. ACL.

[7] M. Elsner, E. Charniak, and M. Johnson. Structured generative models for unsupervised named-entity clustering. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. ACL, 2009.

[8] P. Ferragina and U. Scaiella. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of CIKM*. ACM, 2010.

[9] T. Guardian. The Sarah Palin emails, 2011.

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

[11] T. Lin, Mausam, and O. e. a. Etzioni. No noun phrase left behind: detecting and typing unlinkable entities. In *Proceedings of the 2012 Joint Conference on EMNLP and CoNLL*. ACL, 2012.

[12] X. Ling and D. S. Weld. Fine-grained entity recognition. In *AAAI*, 2012.

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

[14] R. Mihalcea and A. Csomai. Wikify!: Linking Documents to Encyclopedic Knowledge. In *Proceedings of the 16th CIKM*, CIKM '07. ACM, 2007.

[15] D. Milne and I. H. Witten. Learning to link with wikipedia. In *Proceedings of CIKM*. ACM, 2008.

[16] D. Nadeau. Semi-supervised named entity recognition: learning to recognize 100 entity types with little supervision (Ph.D. thesis). 2007.

[17] T. New York Times. The Palin E-Mails, 2011.

[18] Radicati Group Inc. Email Statistics Report, 2010-2014.

[19] L. Ratinov, D. Roth, D. Downey, and M. Anderson. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the ACL*. ACL, 2011.

[20] E. Riloff, R. Jones, et al. Learning dictionaries for information extraction by multi-level bootstrapping. In *AAAI/IAAI*, 1999.

[21] S. Sekine. On-demand information extraction. In *Proceedings of the COLING/ACL on Main conference poster sessions*. ACL, 2006.

[22] Shaun Usher. *Letters of Note*. `http://www.unbound.co.uk/`, 2012.

[23] Stanford University Libraries. Guide to the Stanford Humanities Lab Records.

[24] Sunlight Foundation. Announcing Sarah's Inbox, 2011.

[25] Susan Thomas. Paradigm Academic Advisory Board Report. *John Rylands University Library, Manchester*, Dec. 12, 2005.

[26] F. B. Viégas, S. Golder, and J. Donath. Visualizing email content: portraying relationships from

Mr. Creeley, First of all I want to introduce myself by saying that I am enrolled in the MFA program at **San Diego State**.
Some background: Mack is a grad student at **Johns Hopkins**, where ...
... currently a visiting scholar at univeristy : **Vanderbilt Univ** ...
Ph.D., **University of Southern California**.
The Center for the Humanities **CUNY Graduate Center**

Table 5: Universities

The campaign will be overseen by the Office of **Global Communications**, and will use advertising techniques
C 2003 **Independent Digital (UK) Ltd** Debra Kolodczak
He has also acted as a multimedia consultant for clients like Warner Bros., **Sony Pictures**, Lions Gate Films, New Line Cinema, and Miramax.
The CD is produced by The **Media Workshop**, Inc., but I couln't get any fix on it..
Recorded at **XM Studios** in Washington, D.C., 2003.

Table 6: Companies

Barbara Bono, Shakespeare, **The Cult of Elizabeth and the Productionof Elizabethan Literature**.
The Fate of Stories is about how our place in the world is defined and changed by the stories we read ...
As you know, we did a collaboration called **FAMOUS LAST WORDS**
Another example of an intrusion of the ghost of the past is in the poem **The Rag**
His second novel, **Indian Killer**, published in 1996

Table 7: Books

Thursday, February 7th, 8 p.m. Hallwalls **Contemporary Arts Center**
She was involved in founding **Tokyo Metropolitan Museum of Photography and NTT/ICC**, and is a co-founder of Digital Image
NYC last weekend for the celebration of the **Rubin Museum of Tibetan Art** – any
featuring the collection of the **Museum of International Folk Art**
which hangs in the Scottish **National Gallery of Modern Art** in Edinburgh

Table 8: Museums

he received the **Pulitzer Prize for Poetry**, the **National Book Critics Circle Award**, and the **National Book Award**.
National Academy of Sciences and won the National Book Award and the **National Book Critics Circle Award**

Table 9: Awards

Movie Received **My Architect: A Son's Journey**
preparing a presentation of **The Mystical Romance of Layla & Majnun**
narration for the film **Lowell Blues** was GREAT!
is about the release this weekend of **Going Upriver: The Long War of John Kerry**

Table 10: Movies

conversational histories. In *Proceedings of CHI '06*. ACM, 2006.

[27] M. Wright. Why the British Library archived 40,000 emails from poet Wendy Cope. *Wired*, May 10, 2011.