# Knowledge Base Population for Organization Mentions in Email

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

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

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

## Abstract

A prior study found that on average there are 6.3 named mentions of organizations found in email messages from the Enron collection, only about half of which could be linked to known entities in Wikipedia (Gao et al., 2014). That suggests a need for collection-specific approaches to entity linking, similar to those have proven successful for person mentions. This paper describes a process for automatically constructing such a collection-specific knowledge base of organization entities for named mentions in Enron. A new public test collection for linking 130 mentions of organizations found in Enron email to either Wikipedia or to this new collection-specific knowledge base is also described. Together, Wikipedia entities plus the new collection-specific knowledge base cover 83% of the 130 organization mentions, a 14% (absolute) improvement over the 69% that could be linked to Wikipedia alone.

## 1 Introduction

The Text Analysis Conference Knowledge Base Population (TAC-KBP) track defines several knowledge base population tasks, including linking mentions to corresponding Knowledge Base (KB) entities (i.e., entity linking), extending the KB with newly discovered entities (i.e., NIL clustering), and discovering attribute values for both known and new entities (i.e., slot filling). When linking mentions of well known entities, general-coverage KBs such as those built from Wikipedia are useful. Prior work has, however, found that few people who are mentioned in the course of informal interactions (specifically, in Enron email) exist in such general-coverage KB's (Gao et al., 2014). That fact resulted in renewed interest in constructing collection-specific KBs, which had been investigated a decade earlier in other contexts (Elsayed and Oard, 2006). That approach proved productive, covering about 80% of all person named mentions found in the Enron collection (Elsayed et al., 2008). Several entity linking systems (some referred to in earlier work as "identity resolution") have been proposed for resolving named mentions of people in email messages to a collection-specific person KB (Minkov et al., 2006; Elsayed et al., 2008; Xu and Oard, 2012; Diehl et al., 2006). The next natural question to explore is whether similar techniques might be used to create collection-specific organization (ORG) KBs, since that same earlier study reported that only about half (53%) of ORG mentions in the Enron collection could be found in Wikipedia (Gao et al., 2014). That is our focus in this paper.

Collection-specific person KBs for email collections (Elsayed and Oard, 2006) are built by taking the set of email addresses found as senders or recipients in the collection as candidate entities for which names are then mined (e.g., from headers, salutations, signatures, or the address itself), finally clustering candidate entities that seem to refer to the same person into a single entity. We can draw on the same insight to create an ORG KB by observing that the domain names found in those email addresses provide a set of candidate ORG entities. From there, however, the approach necessarily diverges because collection-internal evidence for the

organization name associated with each of those entities is limited. It is, however, possible to also turn to external sources (e.g., Web search) for clues, since many organizations would be expected to have a Web presence. Another challenge is that some domain names (e.g., hotmail.com) are generic and thus not useful for resolving ORG mentions. To explore these questions, we have built a collection-specific KB from the domain names in the Enron collection. As we report below, organization names and additional information could be found for three-quarters (75%) of the non-generic domain names in the KB as attributes by using a fully automated process, nearly all of which are correct.

Building such a KB is just a first step; the next question is whether the organizations in that KB were actually mentioned often in the collection. To answer this question, we have built a new test collection containing 130 organization mentions, each of which we have tried to link manually (as ground truth) to Wikipedia and to our new collection-specific KB. Our results indicate that the Wikipedia coverage is somewhat higher than previously reported (69% from our manual linking, which is better than the previously reported 53%), and that the proper referent for an additional 14% of organization mentions can be found only in our new collection-specific organization KB. In total, 83% of all organization mentions can be linked to entities in one, the other, or both KB's.

There are two main contributions in this paper. First, we propose a completely automatic process to populate a collection-specific organization KB from email collection. Additional information extracted from four sources are inserted into the KB as attributes for 75% of the ORG entities. Second, we extend an existing entity linking evaluation collection by linking the organization named mentions detected from sampled Enron email messages to both Wikipedia and our new domain-specific KB. The results show that 60% of the named mentions could be linked to the entities in the collection-specific KB, 14% of which are novel entities compared with Wikipedia. Note that the proposed collection-specific KB is not only an additional linking source for organization named mentions, but also provides connections between the organization and person entities (e.g., a link could be created between the person entity "Kenneth Lay" and organization entity "Enron" through the email address "kenneth.lay@enron.com"), which benefits both knowledge base population and entity linking.

The remainder of the paper is organized as follows: Section 2 outlines our process for generating entities from the domain names found in email addresses. Section 3 then describes how we use four specific sources to identify information about the organizations associated with those domain names, providing accuracy and coverage statistics for each source. We then describe our new entity linking test collection in Section 4. Finally, Section 5 concludes the paper with a few remarks about future work.

## 2 Extracting Candidate Organization Entities

In the CMU Enron email collection (Klimt and Yang, 2004), there are 23,265 unique domain names extracted from the 158,097 unique email addresses in the collection as candidate ORG entities. 22,195 of these domain names have two levels (e.g., davis-bros.com) or three levels (e.g., dmi.maxinc.com). The remaining 1,070 domains have between 4 and 6 levels (e.g., dshs.state.texas.us). 39.5% of these unique domains are associated with at least two different email addresses. The most frequently used domain, "enron.com," is associated with 37,687 different email addresses in the collection, which indicates that "enron" might be the affiliation of as many as 37,687 person entities.

Three steps are applied to regularize the domains and merge identical ORG entities: (1) Domains are lower cased; (2) Domain segments that are not representing affiliations (i.e., main, alert, admin, student, exchange, list) are removed from the domain (e.g., alert.enron.com is recorded as enron.com in the KB); (3) Domains are merged by stripping the last element, which is the top level domain (e.g., enron.com and enron.net both become enron). This results in 23,008 entities in the collection-specific ORG KB with their associated domain variants and email addresses.

Table 1 shows the top 5 domains with the largest number of unique email addresses in the Enron corpus. The collection contains two types of domains: domains specific to the organization of the sender

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

**Table 1:** The most frequently used email domains in the Enron corpus and the number of associated unique email addresses. We divide domains into organization domains, which are unique to a specific organization, and email service providers, which are shared by many organizations.

| Source | Domains | Addresses | Accuracy |
|---|---|---|---|
| Google | 68.4% | 83% | 20/20 |
| Wikipedia | 27.6% | 64% | 15/20 |
| Signature | 0.9% | 26.3% | 20/20 |
| Body | 3.4% | 29.2% | 17/20 |
| Overall | 75.1% | 87.7% | |

**Table 2:** Success rate of extracting organization information from different sources.

(e.g., enron, haas.berkeley, dynegy) (**Organization Domains**), and large email service providers (e.g., aol, hotmail) (**Email Service Providers**).

## 3 Extracting Organization Information

Table 2 shows the results of using four different sources (i.e., Google, Wikipedia, Signature and Body of email message) to extract additional information for the organizations. **Domains** and **Addresses** are the percentages of ORG entities and corresponding unique email addresses that can be associated with additional information through one of these sources. **Accuracy** shows the accuracy for the extracted information by manually judging the correctness on 20 randomly sampled non-generic domains (i.e., those that are not email service providers). The methodologies and results are described as following.

**Google** The domain for each ORG entity is submitted to Google as a search query. If the URL of the top returned webpage contains the domain, the webpage is considered as the organization's website. For example, searching for *bluegate* returns the site `http://www.bluegate.com/` with page title *BLUEGATE - Medical Grade Network*. Both the URL and title of the matched webpage are stored as additional information for the ORG entity. Corresponding webpages are found for 68.4% of the ORG entities covering with 83% of the unique email addresses in Enron email collection. To measure the reliability of the Google source, 20 ORG entities with identified webpages were evaluated by the first author; all webpages were judged to be correct.

**Wikipedia** We extract the URL listed in the *Website* and *External Links* fields of Wikipedia Infoboxes and compare them to the domains for ORG entities. The Wikipedia entity with the longest domain segment match is used as the additional Wikipedia link for the ORG entities. For example, the *website* (`www.haas.berkeley.edu`) of the Wikipedia entity *Haas School of Business* has the longest domain segment match with the ORG entity with domain (haas.berkeley). Therefore, the titles and websites of Wikipedia entities are attached to the corresponding ORG entity in the KB. This method identified matches for 27.6% of the ORG entities covering 64% of the email addresses. Manual judgments on 20 matched ORG entities show that 15 of 20 of the entities are matched to the correct Wikipedia entities. When there is more than one segment in the domain of an ORG entity, it usually represents the hierarchy of the organization (e.g., *store.yahoo* represents *Yahoo Store* in *Yahoo!*). When the Wikipedia entity with longest domain segment match is only a partial match, there were misalignments (the 5 errors in the evaluation set) between the ORG in the KB and the ORG entity in Wikipedia. For example, the Wikipedia entity with longest domain segment match is *Yahoo!* (with *Website* `www.yahoo.com`) which is incorrect for the domain *store.yahoo*.

**Signature** Email signatures often contain the affiliation of the sender. We use the approach of Carvalho and Cohen (2004) to detect the signatures in email messages. Phrases with capital initials in the signature are recognized as potential organization names if there is a 5-gram string match between the domain of the sender's email address and the phrase. For example, *Harvard* and *Harvard Business School Publishing* are all valid organization names for domain *hbsp.harvard*. The frequency of the observed organization name / domain pairs are stored for each

ORG entity. Through the source of signature, we identify names for 0.9% of the ORG entities associated with 26.3% of the email addresses, in which 23.8% of the email addresses are registered under "enron" domain. Manual judgments on 20 randomly sampled Non-NIL ORG entities show that all 20 of the extracted organization names are correct.

**Body** Similar to using the signature, organization names can appear in the body of the email. By using the source of email message body, 3.4% of the ORG entities are attached with additional organization information covering 29.2% of the email addresses. Manual judgments on 20 randomly sampled Non-NIL ORG entities show that 13 of the the extracted organization names contain valid information.

**Analysis of the KB** Overall, Google is the best source for finding additional information for ORG entities. Wikipedia, Body and Signature provide coverage for an additional 6.7% of the entities. For example, the first returned Google search page for domain "infoseek" is "go.com" since the search engine "infoseek" was acquired by The Walt Disney Company and merged with the "go.com" network. Since we only examine the first returned result from Google, we miss this match. However, from Wikipedia we can find this historical information from the page of *Infoseek*. Another example is the domain of "ms1.lga2.nytimes". The first Google returned result is a page containing an email address "siteadm@ms1.lga2.nytimes.com", which doesn't provide domain information. However, the Body and Signature contain the strings "The New York Times Company" and "The New York Times".

In total, we identified information for 75.1% of the unique ORG entities covering 87.7% of the email addresses. 60.5% of the domains are associated with only one email address. If considering only the domains with at least two email addresses, additional information can be extracted for 77.8% of the entities covering 89% of the email addresses.

## 4 Entity Linking Test Collection

We now turn to an evaluation of the impact of our KB augmentation on the task of entity linking. Gao et al. (2014) created a collection of annotated email messages with links to a KB derived from Wikipedia. The collection contained 152 named organization mentions, of which 53% could be resolved to Wikipedia. We extended this work by also linking these mentions to our collection-specific organization KB.

We found that 22 mentions (which in the earlier work had been automatically detected) were not actually organizations. For example, "Rio Bravo IV Project" refers to a project rather than an organization. We therefore removed these 22 invalid mentions. For the 130 valid ORG mentions, 60 (46.1%) of them could be resolved to both Wikipedia and the collection-specific KB (e.g., "Pacific Gas and Electric Company" is an entity in Wikipedia and "pge" exists in the collection-specific KB). 30 mentions (23.1%) could only be resolved to Wikipedia (e.g., "Jacksonville Jaguars" are an American football team), while 18 mentions (13.8%) could only be resolved to an entity in the collection-specific KB (e.g., "PIRA" refers to PIRA Energy Group, with domain name "pira" after stripping the top-level domain). 22 of the mentions (16.9%) could not be linked to either KB ("SonoSite").

## 5 Conclusion

We have described a method for automatically populating organization information in a KB based on an email corpus. We gather information from Web sources (Google and Wikipedia) as well as the email collection (body and signature). Our methods identify organization information for 75% of the email domains, covering 87.7% of the unique email addresses in the collection. We show the value of the resulting collection by determining the coverage it provides to an email entity linking task.

Our methods were unable to provide information for one quarter of the entities. We believe additional coverage could be achieved through better processing of domains, such as identifying those that are originators of spam. We also plan to consider additional sources of information. Additionally, our work on organizations could be applied to other publicly available email collections (LDC, 2015), integrated with research into creating person KBs, and evaluated using an end-to-end entity linking system with both Wikipedia and a collection specific KB.

## Acknowledgement

We would like to thank Dr. James Mayfield, Dr. Paul McNamee, Dr. Tim Finin, and Dr. Dawn Lawrie from Human Language Technology Center of Excellence (HLTCOE) at the Johns Hopkins University for their assistance and comments that greatly improved this proposed work.

## References

Vitor R Carvalho and William W Cohen. 2004. Learning to extract signature and reply lines from email. In *CEAS*.

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

Tamer Elsayed and Douglas W Oard. 2006. Modeling identity in archival collections of email: A preliminary study. In *CEAS*, pages 95–103.

Tamer Elsayed, Douglas W Oard, and Galileo Namata. 2008. Resolving personal names in email using context expansion. In *ACL*, pages 941–949.

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

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

LDC. 2015. Avocado research email collection, https://catalog.ldc.upenn.edu/LDC2015T03.

Einat Minkov, William W Cohen, and Andrew Y Ng. 2006. Contextual search and name disambiguation in email using graphs. In *SIGIR*, pages 27–34. ACM.

Tan Xu and Douglas W Oard. 2012. Exploring example-based person search in email. In *SIGIR*, pages 1067–1068. ACM.