Exploring Example-Based Person Search in Email

Tan Xu[1,3] and Douglas W. Oard[1,2,3]
[1]College of Information Studies/[2]UMIACS, University of Maryland, College Park
[3]Human Language Technology Center of Excellence, Johns Hopkins University
{tanx|oard}@umd.edu

ABSTRACT

This paper describes an entity ranking model for example-based person search in email. Evaluation by comparison to manually resolved named references in Enron email yield results that correspond to typically placing the correct entity in the first or second rank.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation, Measurement

Keywords

Entity retrieval, name resolution, email

1. INTRODUCTION

The usual formulation of the *person search* task requires finding a unique page (e.g., the principal home page) for a named person, where the name is provided in isolation as the query [1, 2], but other formulations are possible. In this paper we focus on resolving named mentions in context, a task that we refer to as *example-based person search*. We model our task in a manner similar to the entity linking task in the Text Analysis Conference (TAC),[1] but with a specific mention in an email messages as the "query," and automatically constructed entity models as the items to be ranked. Specifically, given an email message d where a person's name m is mentioned, and a collection of person entities E, find the entity $e \in E$ to which m refers. We report Mean Reciprocal Rank (MRR) rather than TAC's modified B-Cubed score as our principal evaluation measure because our focus is on formative rather than summative evaluation (thus preferring a ranked measure) and because in our present work we test only with mentions that human annotators are able to resolve (obviating the need to score what TAC calls "NILs").

2. A PROBABILISTIC RANKING MODEL

Our task is to rank all $e \in E$ in decreasing order of the probability of e being the correct resolution given m and d: $P(e \mid m, d)$. By applying the chain rule, this can be inferred as in Formula 1:

$$P(e \mid m, d) = \frac{P(e, m, d)}{P(m, d)} \propto P(d) \cdot P(e \mid d) \cdot P(m \mid e, d) \quad (1)$$

where the prior probability of an email message $P(d)$ can be assumed to be uniform, and $P(e \mid d)$ is the probability of entity e being mentioned in email message d. There are several types of complementary evidence that could help us to estimate $P(e \mid d)$, suggesting a log-linear model:

$$P(e \mid d; \lambda) \propto \exp \sum_{k=1}^{K} \lambda_k f_k(e, d) \quad (2)$$

where, $f_k(e, d)$ is one of K ways of estimating $P(e \mid d)$ and λ_k is some weight for $f_k(e, d)$. In this paper we allocate uniform weights to all λ_k, but in general these weights can and should be learned from training data.

To compute $P(m \mid e, d)$, we simplify it as $P(m \mid e)$ by assuming that the probability of seeing some form of mention for an entity will not be affected by its surrounding email. Of course, this may not be true. For example, when writing to a family member, people might refer to another family member using a nickname that they would not normally use to refer to that same person in an email sent to a business colleague. We leave accounting for that factor to future work.

We estimate $P(m \mid e)$ from normalized occurrence counts in the entity model $C'(l_m)$, which count the number of times each lexical form is observed in a header, salutation or signature. We use the Dice coefficient $s(l_m, m)$ to account for partial string matches between the actual mention m and canonical lexical forms l_m:

$$P(m \mid e) := \sum_{l_m \in e} C'(l_m) \cdot s(l_m, m) \quad (3)$$

3. EXPERIMENTS

We take the Enron email collection [5] hosted by CMU[2] as our email collection; this collection contains messages but no

[1] http://nlp.cs.qc.cuny.edu/kbp/2011/

[2] `http://www.cs.cmu.edu/∼enron/`

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
Copyright 2012 ACM 978-1-4503-1472-5/12/08 ...$10.00.

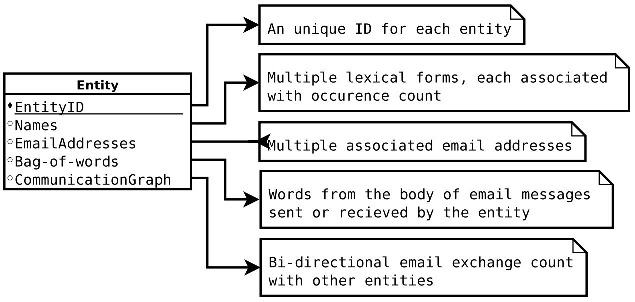
Entity
♦EntityID
◦Names
◦EmailAddresses
◦Bag-of-words
◦CommunicationGraph
An unique ID for each entity
Multiple lexical forms, each associated with occurence count
Multiple associated email addresses
Words from the body of email messages sent or recieved by the entity
Bi-directional email exchange count with other entities

Figure 1: Data Structure for Entity e.

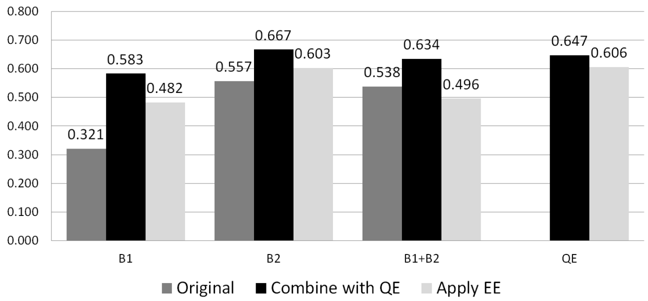

Figure 2: Evaluation Results (MRR)

attachments. We obtained the identity models built automatically for this collection by Elsayed and Oard [4], which includes 123,773 unique entities. Figure 1 shows the data structure for each entity. The bag-of-words representation is stored separately for words in the body of messages sent, received-as-*to* by the entity, and received-as-*cc* by the entity (with quoted text removed). The communication graph contains a list of all entities e' with whom entity e exchanged at least one message in each direction (i.e., at least one message sent as "to" or "cc" to e' and at least one message received as "to" or "cc" from e').

As ground truth we use a publicly available set of 470 single-token mentions (in 285 unique messages) that have been manually resolved to email addresses.[3] The reported inter-annotator agreement is 81% and the median lexical ambiguity (i.e., the number of identity models with an exact string match) is 116 [3].

We tried four ways of estimating $f_k(e, d)$. Method (*B1*), based on Balog's "model 1" [1], finds entities who typically use similar words, regardless of whether those words are used together in one email. The words in the body of d are taken as query terms (with no removal of quoted text), and Lucene's standard analyzer and index searcher are used to score each entity. Method (*B2*), based on Balog's "model 2" [1] works similarly, but indexing individual messages rather than the concatenation of all messages sent or received by each entity. For the top-scoring 50 messages d_r retrieved for d_q, each entity in the "from", "to" or "cc" field of d_r receives an equal portion of d_r's retrieval score and these partial scores are then summed over all 50 documents. Method (*QE*) directly uses d_q's "from" "to" and "cc" fields, assigning each such entity e equal weight. These entities are later used in the communication graph to retrieve their one-hop neighbors e', with each valued by their normalized number of messages exchanged with e. Using this same communication graph, Method (*EE*) is constructed by projecting weights to entities that are directly connected by any ranked entity collected from the three previous methods. Each method's results are first renormalized to sum to 1, and then combined as described in equation (3). Contributions to the same entity are summed, and the final results are renormalized to sum to 1.

4. RESULTS AND CONCLUSIONS

As Figure 2 shows, combining Methods (*B2*) and (*QE*) does rather well, achieving an MRR of 0.667. Since our ultimate goal is to use this as the first stage of an automated entity resolution system (which will also leverage intra-document consistency evidence in subsequent iterative steps), we interpret this as a promising result. Our processing requires an average of 0.78 seconds per query (one core, 2.66 GHz CPU, 4 GB RAM), which meets our efficiency requirements for a preprocessing stage. Elsayed, using different techniques, similarly found the communication graph to be the best source of evidence (MRR=0.785), but at the cost of the substantially greater computational cost to simultaneously resolve all 1.3 million named references to people in the Enron collection [3].

Comparing Methods (*B1*) and (*B2*), our results parallel those Balog et al report for other content types, with (*B2*) markedly better than (*B1*) [1]. Combining evidence from Methods (*B1*) and (*B2*) yields no improvement over Method (*B2*) alone, suggesting that further work on combination methods is called for. Zaragoza et al found the entity graph to be useful for ranking relevant entities [6]; our failure to replicate that by Method (*EE*) suggests that we should exlore graph-based entity ranking measures and that learned weights are likely needed for a log-linear combination in this application. We plan to explore those ideas in future work.

5. ACKNOWLEDGMENTS

We appreciate the advice and suggestions of Tim Finin, Tim Oates, Dawn Lawrie, Paul McNamee, Jim Mayfield, Ves Stoyanov and William Webber.

6. REFERENCES

[1] K. Balog, L. Azzopardi, and M. de Rijke. A language modeling framework for expert finding. *Information Processing & Management*, 45(1):1–19, 2009.

[2] A. P. de Vries et al. Overview of the INEX 2007 entity ranking track. In *INEX*, pages 245–251. 2007.

[3] T. Elsayed. *Identity resolution in email collections.* PhD thesis, University of Maryland, 2009.

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

[5] B. Klimt and Y. Yang. The Enron corpus: A new dataset for email classification research. In *ECML*, pages 217–226. 2004.

[6] H. Zaragoza, H. Rode, P. Mika, J. Atserias, M. Ciaramita, and G. Attardi. Ranking very many typed entities on wikipedia. In *CIKM*, pages 1015–1018. 2007.

[3] `http://www.cs.umd.edu/~telsayed/research.htm`