A Test Collection for Email Entity Linking

Ning Gao, Douglas W. Oard
College of Information Studies/UMIACS
University of Maryland, College Park
{ninggao,oard}@umd.edu

Mark Dredze
Human Language Technology Center of Excellence
Johns Hopkins University
mdredze@cs.jhu.edu

Abstract

Most prior work on entity linking has focused on linking name mentions found in third-person communication (e.g., news) to broad-coverage knowledge bases (e.g., Wikipedia). A restricted form of domain-specific entity linking has, however, been tried with email, linking mentions of people to specific email addresses. This paper introduces a new test collection for the task of linking mentions of people, organizations, and locations to Wikipedia. Annotation of 200 randomly selected entities of each type from the Enron email collection indicates that domain-specific knowledge bases are indeed required to get good coverage of people and organizations, but that Wikipedia provides good (93%) coverage for the named mentions of locations in the Enron collection. Furthermore, experiments with an existing entity linking system indicate that the absence of a suitable referent in Wikipedia can easily be recognized by automated systems, with NIL precision (i.e., correct detection of the absence of a suitable referent) above 90% for all three entity types.

1 Introduction

Connecting content found in free text to structured knowledge sources is a useful step for information access, question answering and knowledge base population. Given a document, named entity recognition (NER) first identifies the names mentioned in the document. Given a knowledge base (KB) and the recognized name mentions, entity linking, also known as named entity disambiguation (NED), links the name mentions to the referent entities in the KB, or returns NIL if the references are not in the KB. With the rise of large scale KBs such as Wikipedia, Freebase and Yago, entity linking has drawn considerable attention. The Text Analysis Conference Knowledge Base Population (TAC-KBP) track introduced the first shared entity linking task [1] in 2009.

Since their introduction, entity linking studies have explored a variety of data types and settings. Traditionally, many systems have sought to link to a KB derived from Wikipedia infoboxes [1, 2]. More recently, Freebase and Yago have also been used as even more extensive sources of structured knowledge [3, 4]. Some shared tasks, such as TAC-KBP, have focused only on linking named entities, but others have extended the task to link mentions of any concept found in Wikipedia, so called “Wikification” [5, 6]. The TAC-KBP shared task has also evolved, adding NIL clustering (clustering of entities not found in the KB) [7, 8], extending from an initial focus on linking from news articles to also include linking from blogs [8], and adding tasks that required linking across languages [7, 9]. Furthermore, several recent papers have considered social media, such as Twitter, as a new source for mentions to be linked to entities [10, 11, 12, 13].

There has been another thread of work on identity resolution in email, which we can think of as a specialized entity linking task. The focus of that work has been on automatically tagging named mentions of a person in the body text of email message with the email address of that person. If we think of the email address inventory as a domain-specific KB, then this is an entity linking task. However, identity resolution research has to date focused only on person entities, leaving scope for

Table 1: Statistics for the sampling of named mentions, and the NER accuracy

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

Table 2: Wikipedia coverage statistics.

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

Table 3: Entity linking accuracy, HLTCOE Entity Linker.

Type	Non-NIL	Exact	NIL	Overall
PER	0.517	36%	0.934	0.799
ORG	0.688	39%	0.944	0.809
LOC	0.729	40%	0.923	0.794

future work on organizations, locations, and other entity types. Moreover, identity resolution research has to date focused only on linking to entities that are uniquely identified by email addresses, leaving scope for future work on resolving mentions of people for which no suitable referent email address is known (e.g., mentions of public figures).

As an initial step towards addressing these challenges, we have developed what we believe to be the first test collection for linking named mentions of person (PER), organization (ORG) and location (LOC) entities from email message bodies to Wikipedia.[1] As a second step, we have used the existing HLTCOE entity linker [14] to establish a baseline for entity linking accuracy to which future more highly specialized systems can be compared. In the remainder of this paper, we first describe our new test collection in Section 2. We then report the entity linking accuracy of the HLTCOE entity linker on the new collection in Section 3. In Section 4, we recap some related work. Section 5 concludes the paper with a brief discussion of next steps.

2 Test Collection

We use a de-duplicated Enron email collection as the basis for our new test collection. Within the CMU Enron email collection [15], there are a substantial number of duplicate email messages (because the same email could, for example, appear in the sender's Sent Mail folder and the recipient's Inbox folder, and because some users keep multiple copies of email messages, so that the same message might be in the Inbox and also in a folder called "East Oil"). We therefore adopted Elsayed's de-duplication process [16], in which email messages are considered to be duplicates if they contain exactly the same From, To, Cc, Bcc, Subject, Time, and Body fields. Before de-duplication we normalized the date and time of each message to a standard time zone (Universal Coordinated Time). After de-duplication there are 248,573 email messages in the collection.

To automatically identify named mentions (i.e., omitting nominal and pronominal mentions), we first use the Illinois Named Entity Tagger (INET)[2] to recognize three categories of named entity mentions: person (PER), organization (ORG), and location (LOC). In Table 1, the column labeled **All Msg** shows the number of named mentions recognized by INET in the whole collection, with the **Per Msg**$_{\textbf{All}}$ column showing the average number of mentions in each email. We then randomly selected 300 messages that each contained more than 10 words of body text. The **Per Msg**$_{\textbf{300}}$ column shows the average number of mentions in the sampled 300 messages.[3] Finally, from the detected entities in those 300 messages (shown in the **300 Msg** column) we then randomly selected 200 named mentions (**Sample** size) of each type. Comparing the numbers of **Per Msg**$_{\textbf{All}}$ and **Per Msg**$_{\textbf{300}}$, we can see that the prevalence of named entity mentions is somewhat higher in the sampled collection than in the whole collection because the statistics for the whole collection are reduced by the presence of short messages that contain relatively few named mentions.

[1] The test collection is available at `http://www.umiacs.umd.edu/~ninggao/publications`

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

[3] We randomly select 20 documents and manually recognize 135 PER, ORG or LOC mentions. The estimated recall of INET on the sampled documents is 98.5%.

Six annotators then each labeled 100 of the 600 sampled entity mentions for whether the text span recognized by INET was a correctly delimited named mention of an entity of the corresponding type. For example, in the sentence "I will meet him in Washington DC," the only correct text span would be "Washington DC" (not "Washington" or "in Washington DC"), and "Washington DC" should be classified by INET as a LOC, not a PER. Because each named mention was judged for correctness by a single annotator, measures of inter-annotator agreement for this task are not yet available. As the **Correct** and **Accuracy** columns in Table 1 show, the accuracy for PER and LOC mention detection is comparable to levels typically achieved on newswire (90% and 97%, respectively), but detection accuracy for ORG mentions is considerably lower (76%). For the 10% of PER mentions that were missed, informal writing styles in which plausible (but incorrect) names such as "Hope" or "May" can begin a sentence make the task more challenging, particular for systems like INET that are not trained specifically for email. As expected, we also see classes of errors that are also typical in newswire, such as incorrectly segmenting "George Bush" from "George Bush Intercontinental Airport." The few errors on LOC all resulted from typical causes that are also seen in newswire (e.g., misrecognizing "Turkey" as a LOC when, read in context, it is clearly referring to a bird). In the 24% of automatically detected ORG mentions marked by our assessors as incorrect, we find product names mislabeled as ORG (e.g., "MAC") and job positions mislabeled as ORG (e.g., "ITC" in "if you want to be an ITC"). We also see some domain-specific names that our assessors simply lacked the knowledge to to judge with confidence (e.g., "RTO" in "standardizing RTO.")

To measure the accuracy of automated entity linking systems, we also asked each assessor to provide us with the correct Wikipedia page for each correctly recognized mention that they assessed, or NIL to indicate if they believed that no such Wikipedia page yet existed. The **Non-NIL** and **NIL** columns in Table 2 show the number of entities that the annotator had judged as correct that were or were not in Wikipedia, respectively. A sample of 60 these mentions (20 of each type) was dual annotated by a second assessor, yielding exact agreement (i.e., both designate the same entity or both designate NIL) on 85% of the named mentions. The Cohen's Kappa agreement on whether a mention was NIL or Non-NIL is 0.933. From Table 2, we can see that only 32% of the PER mentions could be linked to Wikipedia entities. These PER entities include sport stars, politicians, and well known people who worked for Enron such as the former CEO Kenneth Lay. For ORG mentions, about half (53%) of the referenced entities were found in Wikipedia (e.g., "Justice Department") with the other half annotated as NIL (e.g., "Southward Energy Ltd"). Most (93%) of the LOC mentions could be found in Wikipedia; the relatively few LOC mentions that were resolved by our annotators as NIL included references to specific locations that had not achieved sufficient notoriety for inclusion in Wikipedia (e.g., "1455 Pennsylvania Ave").

3 Entity Linking Result on the New Collection

We then ran an existing entity linking system, the HLTCOE entity linker [14], on the 524 correctly recognized name mentions that had been automatically found by INET. Table 3 shows the resulting accuracy for mentions that had been resolved by our annotators to Wikipedia (**Non-NIL**) and for mentions that had been marked by our annotators as **NIL**. Aggregate accuracy statistics (**Overall**) are also shown. The **Exact** column in Table 3 shows the fraction of the correctly resolved **Non-NIL** mentions that resulted from an exact match to the canonical form in Wikipedia (which is unique); the other correct resolutions resulted from finding candidates using a fuzzy string match, ranking those candidates using match degree, entity popularity and contextual similarity features, and then selecting the top-ranking match.

Our inspection of the results indicates that the exact matches were always correct, but that the difficulty of fuzzy matching varied by entity type. For PER mentions, the entity linker tends to resolve single-token name mentions, such as "Parker" in "Parker could get a touchdown pass Sunday against Seattle Seahawks," to NIL when there's not enough useful evidence available to the entity linker. However, our human assessor could resolve the mention to the American football player "Larry Parker" after some easy reasoning. A frequently observed fuzzy match pattern for ORG name mentions that the linker often got right was abbreviation expansion (e.g., matching a mention of the "US congress" to "United States Congress"). Only for LOC mentions were more mentions correctly resolved through fuzzy matching than through exact match match. This reflects the rarity of fully qualified location references in Wikipedia (e.g., we see what Wikipedia calls "Orlando, Florida" mentioned simply as "Orlando").

Looking at Tables 2 and 3 together, we can conclude that Wikipedia KB linking achieves good coverage (a true non-NIL rate of 93%) with fairly high accuracy (0.73, despite the frequent need for fuzzy matching) for LOC entities. The situation for PER entities is just the opposite, with Wikipedia KB linking achieving quite poor coverage (with a true non-NIL rate of 32%) and with automated entity linking doing somewhat less well for those cases (with an accuracy of 0.52, about half of which are exact matches on a full name). The situation for ORG entities is somewhere in between, with a substantial number of true non-NIL links to the Wikipedia KB that are possible (53%), but rather low linking accuracy (0.69). From this we conclude that the research to date on linking PER entities to domain-specific KB's has been well justified, that new research on linking ORG entitles to domain-specific KB's is called for, and that research on linking LOC entities to domain-specific KB's can reasonably continue to be given lower priority. Notably, Table 3 also shows that the HLTCOE entity linker is already rather good at NIL detection on mentions found in email messages, with detection accuracy above 0.9 for true NIL's on all three types, despite not having been tuned specifically to email. This result suggests that it should be fairly easy to integrate the results of entity linking to broad-coverage and domain-specific KB's.

4 Related Work

Klimt and Yang [15] published the CMU Enron email collection, built from 152 users' email folders, totaling 517,424 messages (but no attachments). Minkov et al. [17] built the first evaluation collection from Enron emails using messages from the folders of two users (Sager and Shapiro) that contained named mentions that corresponded uniquely to names in the CC field. The corresponding names were removed from the CC field for evaluation purposes. Diehl et al. [18] built a test collection with 78 named mentions that were manually annotated as resolving to specific Enron email addresses. Elsayed [16] built what we can think of as the first collection-specific KB for the Enron email collection, containing not just the email address, but also known name variants and known alternate email addresses for the same person. Elsayed also built what is currently the largest manually annotated test collection for the task of person entity linking in Enron email.

McNamee et al. [14] created the HLTCOE Entity Linker, which we used in Section 3. The system first computes a set of triage features (e.g., string equality, approximate string match) and selects candidate KB entity based on those features. NIL is then added to every candidate set, and ranked in the same way as any other candidate. An extended set of more expensive features (e.g., document similarity, popularity, and plausible NIL cues) are then computed for each candidate. Finally, the candidates are ranked using those features and a learned ranking function, and the top-ranked candidate is returned as the system's prediction of the referent entity. The HLTCOE Entity Linker has proven to be one of the most effective systems in the TAC-KBP entity linking task[1].

5 Conclusion and Future Work

From Table 2, we can see that about two-thirds of the mentioned person entities and about half of the mentioned organization entities in the Enron email collection are not covered by Wikipedia. It is, therefore, potentially useful to build collection-specific KBs for those entity types; location entities (for which only about 7% are missing from Wikipedia) seem to us currently to be less of a priority. Although Elsayed et al. [19] have produced a collection-specific KB for persons who sent or received messages in the Enron email collection, no comparable collection-specific KB yet exists for organizations. We therefore believe that research on that topic deserves attention. Further, the NIL mentions in the proposed test collection could be resolved to the collection-specific KBs. We expect that the linking to collection-specific KBs will largely reduce the percentage of NIL mentions. We are also interested in the clustering for the rest of the NIL mentions.

Our current test collection only includes three entity types: PER, ORG and LOC. In our future work, we are interested in extending the entities to other types such as events and products. We also note that the NER and entity linking systems that we have used were not designed to exploit characteristics of email that could make the task easier (e.g., thread structure, the communicant graph, or bursty communication patterns). Leveraging such features could be important because in some ways entity linking in email is harder than entity linking in news. For example, news articles might be expected to be explicitly self-contextualizing more often that isolated email messages in which individual senders and recipients rely more heavily on shared context.

Finally, we note that forcing a choice between broad-coverage and collection-specific KB's would likely be suboptimal. We would expect any collection to contain references to well known and less well known entities, so developing techniques for linking from one collection to multiple KBs deserves attention from researchers. Email provides a particularly attractive testbed around which to develop such capabilities since, as our work has shown, both types of references are naturally present in substantial quantities. In this way, research on entity linking from email collections can help to develop capabilities from which the field as a whole can ultimately benefit.

References

[1] P. McNamee and H. T. Dang, "Overview of the TAC 2009 knowledge base population track," in *TAC*, 2009.

[2] S. Cucerzan, "Large-scale named entity disambiguation based on Wikipedia data.," in *EMNLP-CoNLL*, pp. 708–716, 2007.

[3] Z. Zheng, X. Si, F. Li, E. Y. Chang, and X. Zhu, "Entity disambiguation with Freebase," in *IEEE/WIC/ACM*, pp. 82–89, 2012.

[4] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum, "Robust disambiguation of named entities in text," in *EMNLP*, pp. 782–792, 2011.

[5] D. N. Milne and I. H. Witten, "Learning to link with Wikipedia," in *CIKM*, pp. 509–518, 2008.

[6] L. Ratinov, D. Roth, D. Downey, and M. Anderson, "Local and global algorithms for disambiguation to Wikipedia," in *ACL*, 2011.

[7] H. Ji, R. Grishman, and H. Dang, "Overview of the TAC 2011 knowledge base population track," in *TAC*, 2011.

[8] H. Ji, R. Grishman, H. T. Dang, K. Griffitt, and J. Ellis, "Overview of the TAC 2010 knowledge base population track," in *TAC*, 2010.

[9] P. McNamee, J. Mayfield, D. W. Oard, T. Xu, K. Wu, V. Stoyanov, and D. Doerman, "Cross-language entity linking in Maryland during a hurricane," in *TAC*, 2011.

[10] T. Cassidy, H. Ji, L.-A. Ratinov, A. Zubiaga, and H. Huang, "Analysis and enhancement of Wikification for microblogs with context expansion.," in *COLING*, pp. 441–456, 2012.

[11] S. Guo, M.-W. Chang, and E. Kiciman, "To link or not to link? a study on end-to-end tweet entity linking.," in *HLT-NAACL*, pp. 1020–1030, 2013.

[12] X. Liu, Y. Li, H. Wu, M. Zhou, F. Wei, and Y. Lu, "Entity linking for Tweets," in *ACL*, pp. 1304–1311, 2013.

[13] W. Shen, J. Wang, P. Luo, and M. Wang, "Linking named entities in Tweets with knowledge base via user interest modeling," in *SIGKDD*, pp. 68–76, 2013.

[14] P. Mcnamee, M. Dredze, A. Gerber, N. Garera, T. Finin, J. Mayfield, C. Piatko, D. Rao, D. Yarowsky, and M. Dreyer, "HLTCOE approaches to knowledge base population at TAC 2009," in *TAC*, 2009.

[15] B. Klimt and Y. Yang, "The Enron corpus: A new dataset for email classification research," in *ECML*, pp. 217–226, 2004.

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

[17] E. Minkov, W. W. Cohen, and A. Y. Ng, "Contextual earch and name disambiguation in email using graphs," in *SIGIR*, pp. 27–34, 2006.

[18] C. P. Diehl, L. Getoor, and G. Namata, "Name reference resolution in organizational email archives.," in *SIAM*, pp. 70–91, 2006.

[19] T. Elsayed and D. W. Oard, "Modeling identity in archival collections of email: A preliminary study.," in *CEAS*, pp. 95–103, 2006.