Support for Interactive Identification of Mentioned Entities in Conversational Speech

Ning Gao  
University of Maryland  
College Park, MD 20742  
ninggao@umd.edu

Douglas W. Oard  
University of Maryland  
College Park, MD 20742  
oard@umd.edu

Mark Dredze  
The Johns Hopkins University  
Baltimore, MD 21218  
mdredze@cs.jhu.edu

ABSTRACT
Searching conversational speech poses several new challenges, among which is how the searcher will make sense of what they find. This paper describes our initial experiments with a freely available collection of Enron telephone conversations. Our goal is to help the user make sense of search results by finding information about mentioned people, places and organizations. Because automated entity recognition is not yet sufficiently accurate on conversational telephone speech, we ask the user to transcribe just the name, and to indicate where in the recording it was heard. We then seek to link that mention to other mentions of the same entity in a variety of sources (in our experiments, in email and in Wikipedia). We cast this as an entity linking problem, and achieve promising results by utilizing social network features to help compensate for the limited accuracy of automatic transcription for this challenging content.

CCS CONCEPTS
• Information systems → Task models;

KEYWORDS
Speech Retrieval, Entity Linking, Knowledge Base

ACM Reference format:
DOI: http://dx.doi.org/10.1145/3077136.3080688

1 INTRODUCTION
Speech retrieval has been a topic of long standing interest [8]. Early work on speech retrieval focused on formal speech found in news broadcasts, political speeches, and classroom lectures, in part because the accuracy of the Automatic Speech Recognition (ASR) systems used to generate the searched text benefited from the clear articulation, limited vocabulary and formal grammar that is characteristic of formal speech. More recently, however, fairly good retrieval results have been demonstrated for conversational speech as well [12]. At the same time, it is becoming increasingly easy to create large collections of conversational speech. For example, nearly every teleconferencing service provides such capabilities, some lifelogging technologies can (when permitted by law) capture conversational speech easily, "talk shows" with debating panelists have become a pervasive element of the media landscape, and video sharing platforms containing a multitude of types of speech have become ubiquitous.

The question thus arises: what happens after a speech retrieval system has presented some conversational speech to the user? One characteristic of conversational speech is that conversation participants fluidly use references that make sense to them, but may be unclear to a person who later encounters that recording, out of context, as the result of a search. For example, the collection of 1,731 freely available telephone recordings used in this paper were made by Enron employees who were engaged in regulated energy trading activities. References to "Reliant," "Four Corners," or "Jim" that made sense to conversation partners at the time might be completely opaque to a later searcher who finds a call containing those mentions.

Our solution is to provide entity context via a knowledge base (KB), and the goal of this paper is to link specific name references to one or more KBs that can provide additional information about the mentioned entity. Our initial attempt to do this using classic text-based techniques failed miserably because accurately transcribing uncommon proper names is among the most challenging tasks for an ASR system, and as a result the name was often not found in the text. This led us to shift to a more practical approach in which a searcher, listening to a retrieved recording, simply types the name they heard spoken and indicates to the system where in the recording that name was heard.1 We cast this as an entity linking task and simulate the process by selecting mentions to be resolved from 61 manually transcribed recordings. To perform the linking we use only the mentioned name, the point in the recording where that name occurred, and data that can be generated directly from speech using fully automated processing.

The entity linking task has been widely studied for dissemination-oriented media (e.g., broadcast news), and for dissemination-oriented social media (e.g., Twitter) in which the principal focus is on messages meant for broad distribution. Entity linking for conversational media offers new challenges. Some references (31% in our collection) are to well-known entities, while others (69%) refer to entities that would be unlikely to appear in any general-purpose KB. Therefore we attempt to link both to a general-purpose Wikipedia KB and to two automatically constructed application-specific KBs—one for people and one for organizations. The participants often rely

1This interaction is similar to how a user might interact with a zero-resource speech retrieval system [13].
on shared context that may not be explicitly stated in the conversation. We address this difficulty by relying in part on social features constructed from the communication graph. Our results indicate that fairly good performance on this task—as measured by Mean Reciprocal Rank (MRR)—is possible (e.g., MRR=0.78), and that social network features are particularly helpful.

2 ENTITY LINKING SYSTEM
We introduce our entity linking system (§2.1) and then focus on feature construction (§2.2), an important aspect of that system.

2.1 System Framework
Figure 1 shows the framework of the proposed entity linking system. We use a standard feature-based supervised linking system [3] composed of four stages: query preparation, candidate triage, feature construction and prediction. In the query preparation stage, for each designated mention \(q_i\), the context of the mention \(q_i := f(E_i, T_i)\) is extracted, where \(E_i\) is a set of entities that participate in the telephone recording, and \(T_i\) is a vector of words representing the transcript’s content. Figure 1 shows three designated mention examples in the query telephone recording, the PER mention “John”, the ORG mention “Pacificorps”, and the Geo-Political Entity (GPE) mention “Albania”. The participants in conversation \(E_i\) are Jeremy Taylor, Leaf Martin and Holden Johnson. The candidate triage step identifies possible candidates from all the three available KBs for each designated mention based on a cascade of standard heuristics: (1) exact string match; (2) match on initials (e.g., entity Imperial Irrigation District is a match for mention IID); (3) fuzzy match to check if the entity name contains all the words in the mention, or the mention contains all the words in the entity name (e.g., entity United States of America is a match for mention United States); (4) character 4-gram string match; (5) \(\phi\) is a candidate for all the queries to indicate that the true referenced entity is absent from all available KBs. The entities identified by these five steps are designated the candidate set \(E_i\) of mention \(q_i\).

The third step generates a large set of features for each (mention, candidate) pair from the triage phase, which are then used to score each candidate for the given mention in the conversation. Our features are organized into four groups for presentation purposes (§2.2.). All types of candidates share the same set of features. We rank candidates with an SVM regression model (nu-SVR with a radial basis kernel) from LibSVM [2]. The top scoring candidate is the system’s prediction, but we also evaluate the ranked list itself.

2.2 Feature Design
Each feature \(D(q_i, e \in E_i)\) used in the system indicates if the candidate \(e \in E_i\) is the true referent for designated mention \(q_i\). Since previous work for entity linking speech considered formal speech [1], we needed to create new features more suitable for conversational speech. All features in the following four groups are used for all the (mention, candidate) pairs.

2.2.1 General features. General features are designed for all types of designated mentions, including features that measure if any of the name variants of the person candidate entity match the name of a known speaker in the same recording

\[
D(q_i, e) := |e \cap E_i|, \quad (1)
\]

and features measuring if there is a string or fuzzy match between the query \(q_i\) and the name variants \(N\) of the candidate entity

\[
D(q_i, e) := |n \in N : n = q_i|, \quad (2)
\]

Lexical features measure the similarity between the transcript of the recording and the context of the candidates

\[
D(q_i, e) := TF \times IDF, \quad (3)
\]

where all the words in the transcript are used as query terms TF, and the IDF of the terms are calculated using the CMU version of the Enron email collection [10]. For the Wikipedia entities, the context is the content of the entity page. For the entities in the PER KB, the context is the set of manually transcribed recordings in which the entity was known to have participated. For the organization entities, the context is all the name variants that can be found in Google, Wikipedia, email bodies and email signature blocks [6]. We also build a feature \(\text{isNIL}\) indicating that the candidate being ranked is the NIL candidate (\(\phi\)). All the features in the General group have been used with success in related work [3, 5, 11], so we use these General features as one baseline in our experiments.

2.2.2 Person-specific features. We adapt the work of person entity linking for email in [7] to build person-specific features, but only for PER KB candidates. For ORG KB or Wikipedia candidates, these features are set to zero. We build person-specific Social Context features to measure the social similarity between the candidate and the speakers. For each entity, we create a contact list by finding all entities that ever communicated with that entity. Using these lists, we add features to detect if the candidate is in the contact lists of the speakers. We also estimate the probability that a speaker mentions the candidate given that speaker mentions any person with the same name variant

\[
D(q_i, e) := \sum_{c_i \in E_i} \frac{C_{e_i, e}}{\sum_{c'_i \in E_i} |C_{e_i, c'_i} : N_j \cap q_i \neq \emptyset|}, \quad (4)
\]

where \(C_{e_i, e}\) is the frequency that the two entities \(e_i\) and \(e\) are observed in the same conversation, and \(E_p\) represents all the entities in the PER KB. We measure how related the candidate is to the communication group using all communications between the speakers as evidence. Assume that the conversation has a speaker set \(P_q\), and each of the previous \(M\) conversations in the collection between the speaker who starts the conversation of \(q_i\) and candidate have speaker sets \(P_m\), then we add features measuring the social context similarity between \(P_q\) and \(P_m\) based on the Jaccard similarity:

\[
S_{q,m} = \frac{|P_q \cap P_m|}{|P_q \cup P_m|} \times F_m, \quad (5)
\]

where \(m \in [1, M]\), and \(F_m\) is the number of conversations between social group \(P_m\). We build features for \(\text{max}_m S_{q,m}\), the mean over \(m\) of \(S_{q,m}\), the sum over \(m\) of \(S_{q,m}\) and the maximum over \(m\) of a variant of \(S_{q,m}\) in which we first binarize \(F_m\).
2.2.3 Organization-specific features. This group of features is designed for only ORG KB candidates (and set to zero for PER KB or Wikipedia candidates). Each entity \( e \) in the collection-specific ORG KB is a triple \( e_o := (D, N, A) \) in which \( D \) is some unique email domain name (e.g., enron.com), \( N \) are the known name variants for that entity, and \( A \) is the set of email addresses that include \( D \). We build features for the number of email addresses that use each domain name \( D(q,e) := |A| \); the number of levels in the domain name for each ORG candidate (e.g., this value for store.yahoo.com is 3); and the lowest level at which there is a string match between the organization domain and designated mention (if any), counting from the left (e.g., there is a string match between a mention Yahoo and an ORG domain name store.yahoo.com at level 2).

2.2.4 KB-specific features. The KB-specific feature group includes features indicating if the current candidate entity is from a collection-specific PER KB \( E_p \), collection-specific ORG KB \( E_o \) or the general KB built from Wikipedia \( E_w \).

\[
\mathcal{D}(q,e) = \begin{cases} 0, & \text{if } e \in (E_p, E_o) \\ 1, & \text{if } e \in E_w \end{cases}.
\]

For candidate entities from Wikipedia, we include the number of in-links for each entity as a feature. To better match the time frame of our Enron recordings, we used the TAC 2009 Wikipedia KB, which was created from a 2008 English Wikipedia dump.\(^5\) Wikipedia has gotten better over time, and 21 of the referenced entities that are absent from the TAC 2009 Wikipedia KB do have entries in more current versions of Wikipedia. To help our classifier recognize such cases as NILs (references that cannot be linked to a known entity), we include a feature to indicate if the mention has an exact match to the name of a current Wikipedia page, although the entity described by the current Wikipedia page is not included in our KB.

3 TEST COLLECTION

We introduce a new test collection to support evaluation. The full collection contains 1,731 audio files, each of which includes one or more calls made by or to Enron traders, which together total 47.8 hours of conversations.\(^9\)\(^7\) The Snohomish County Public Utility District manually created transcripts for 61 of the 1,731 recordings for use in court, from which we manually extract the designated mentions used in our experiments. There are three KBs used as the linking targets: (1) the TAC 2008 KBP Reference Knowledge Base (which contains PER, ORG and GPE entities); (2) a collection-specific PER KB \(^4\) containing 124,475 person entities; and (3) a collection-specific ORG KB \(^6\) containing 23,008 organization entities. In the PER KB, email addresses are extracted to represent person entities, while the name variants of the entities are extracted from the email header, salutations and signatures. The ORG KB is built based on domain names found in the email addresses of senders and recipients of messages in the collection.

The first author of this paper annotated named mentions and KB links (including NIL) in all three KBs for the 540 PER, ORG and GPE mentions in the 61 manual transcripts.\(^4\) For PER and ORG mentions, a referent entity might be present in both the Wikipedia KB and the corresponding collection-specific KB (e.g., Enron). The mentions include misspellings (e.g., Holli misspelled as Holly), abbreviations (e.g., LV Co-gen), and initials (e.g., ISO). Most of the person mentions are first names or nicknames (e.g., Ken). Table 1 summarizes the linking annotations used as ground truth. A second annotator independently linked a randomly selected half of the person name mentions. This yielded an exact match agreement of 0.78 for the cases in which the first annotator had made a link. A third annotator independently linked 20 randomly selected ORG and 20 randomly selected GPE mentions. The agreement with the first author of this paper on the ORG and GPE mentions is 0.85 and 0.90, respectively.

4 EXPERIMENTS

We evaluated the entity linking system on all three mention types. We based our lexical features on errorful automatic transcripts generated by the Microsoft Oxford Speech API.\(^3\) Table 2 shows the MRR of our system using All features on automatic transcripts.

---


\(^{2}\)https://catalog.ldc.upenn.edu/LDC2014T16

\(^{3}\)https://www.projectoxford.ai/speech

\(^{4}\)http://www.umiacs.umd.edu/~ninggao/publications
Since no prior work links entities in conversations to multiple KBs, we construct two baselines ourselves: (1) only our General feature group, and (2) a Random baseline that randomly selects one entity from the triaged candidate set. Reassuringly, our system does much better than the baselines for all three entity types when evaluated on All mentions in the test set.

Using only General features, the linking result for PER is much worse than for ORG or GPE. Unlike in news articles, mentions of people in these conversations are mostly just first names (e.g., John) or nicknames (e.g., Bill), many of which result in hundreds of candidates (an average of 314 in our collection). Moreover, as shown in Table 1, most (90%) of the named mentions of people refer to entities that can only be found in our collection-specific PER KB, and that KB contains less and sparser context than our Wikipedia KB. Human disambiguation of entity mentions in conversational speech relies heavily on shared context, and indeed we observe that by adding Social Context features (§2.2.2), the MRR for PER mentions improves from 0.273 to 0.779.

Many of the errors in linking ORG mentions arise from changes in organization names due to mergers and acquisitions, which change the name of a company. For example, Reliant Energy (one of the ORG mentions) was renamed NRG Energy after the construction of the ORG KB, and that KB contains less and sparser context than our Wikipedia KB. Human disambiguation of entity mentions in conversational speech relies heavily on shared context, and indeed we observe that by adding Social Context features (§2.2.2), the MRR for PER mentions improves from 0.273 to 0.779.

In contrast, errors in linking GPE mentions are typically caused by changes in the location of a place. For example, the same speakers mentioning “Four Corners” in a short conversation might indicate the referent to be “Four Corners, California”. It is (on average) harder for our system to correctly detect NIL references that should not be linked than it is to link Non-NIL references to the correct entity. Considering both NIL and Non-NIL references, our overall MRR for each entity type is in a fairly narrow range between about 0.7 and 0.8, indicating that the correct referent (or NIL) is often found in the first or second position in the ranked list. These results are below scores reported for newswire (MRR above 0.9), but with feature designs that model some of the context available to the participants, we can achieve linking accuracy that could be useful in a practical search system.

### 5 CONCLUSION

This paper focused on supporting users of retrieval systems for conversational speech by building a new test collection to simulate the task of a user “pointing at” a mention and asking “who or what is that?” One possible direction for future work would be to automate the detection of mentions by tailoring spoken term detection techniques. Since there are typically multiple entity mentions in a conversation, and since the referents of those mentions might be related, we are also interested in resolving all the mentions in the same conversation collectively.

### 6 ACKNOWLEDGMENTS

This work has been supported in part by NSF Grants 1065250 and 1618695 and by a Mellon Foundation Coherence at Scale Doctoral Fellowship. Opinions, findings, conclusions or recommendations are those of the authors and do not necessarily reflect the views of NSF or the Mellon Foundation.

### REFERENCES