Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 65–73
Marseille, 11–16 May 2020

A Study on Entity Resolution for Email Conversations

Parag Pravin Dakle, Takshak Desai, Dan I. Moldovan
Department of Computer Science
The University of Texas at Dallas
{paragpravin.dakle, takshak.desai, moldovan}@utdallas.edu

Abstract

This paper investigates the problem of entity resolution for email conversations and presents a seed annotated corpus of email threads labeled with entity coreference chains. Characteristics of email threads concerning reference resolution are first discussed, and then the creation of the corpus and annotation steps are explained. Finally, performance of the current state-of-the-art deep learning models on the seed corpus is evaluated and qualitative error analysis on the predictions obtained is presented.

Keywords: entity resolution, email threads, coreference resolution, enron, email conversations

1. Introduction

Entity resolution has been an active topic in the Natural Language Processing domain since the 1960s. Shared tasks such as CoNLL 2012 (Pradhan et al., 2012) and MUC (Grishman and Sundheim, 1996) define it as linking referring spans of text that point to the same discourse entity. Previous works highlight that good performance on these tasks does not necessarily result in a similar performance on downstream or similar tasks like machine translation (Guillou, 2012) or anaphora resolution (Aktaş et al., 2018) respectively.

Email corpora have been widely used for numerous tasks like text classification (Klimt and Yang, 2004), intent classification (Cohen et al., 2004), searching (Soboroff et al., 2006) and summarization (Ulrich et al., 2008). This paper addresses the task of entity resolution in email conversations, which to the best of our knowledge has not been examined in any study so far.

This research makes the following key contributions:

1. For the first time, the entity resolution task for emails is analyzed. Characteristics of emails that make this problem difficult are identified.

2. A human-annotated seed corpus containing email threads is presented for the entity resolution task. This annotated seed corpus will be released as a part of the paper[1].

3. An evaluation of the current state-of-the-art model for within document (WD) entity coreference task on the seed corpus and qualitative error analysis on the predictions of the same model is presented.

The paper is organized as follows: Section 2 presents an overview of the work done on the entity resolution task. Section 3 describes the entity resolution task for emails, which is followed by a description of the corpus creation process in Section 4. Challenges observed in the seed corpus as well as email conversations in general are elaborated in Section 5. Section 6 describes the experiments performed on the seed corpus. Error analysis on the predictions is covered in Section 7 and a conclusion is provided in Section 8.

2. Related Work

Over the years, numerous corpora have been released for coreference resolution, with MUC-6 (Grishman and Sundheim, 1996), MUC-7 (Chinchor, 1998), ACE (Doddington et al., 2004) and OntoNotes being the popular ones. OntoNotes 2.0 and OntoNotes 5.0 were used in Task-1 of SemEval 2010 (Recasens et al., 2010) and CoNLL 2012 shared task (Pradhan et al., 2012) respectively. However, each of these corpora either fully or mainly comprise of news articles.

Conversational texts in the form of telephonic speech have been a part of many corpora. The first corpus to entirely focus on conversational texts was Character Identification Corpus (Chen and Choi, 2016). This was constructed using TV show transcripts and labeled speakers in a multi-party conversation. It introduced the task of character-linking in a multi-party conversation. Zhou and Choi (2018) further expanded this corpus by annotating additional text and adding plural mentions. The Manually Annotated Sub-Corpus (MASC) (Ide et al., 2008) project is one of the first corpora to consider annotating coreference chains for emails. The corpus includes 45 emails from the Enron Email Corpus (Klimt and Yang, 2004), 96 spam emails and 35 w3c email digests. However, no coreference annotations were released as a part of the corpus. Furthermore, the emails considered from the Enron Corpus were single messages as compared to our work which focuses on email threads. Hendrickx and Hoste (2009) studied coreference resolution in conversations present in a threaded format. A corpus consisting of blogs and commented news was used to highlight a significant performance drop when dealing with coreference resolution in unedited text. Aktaş et al. (2018), in their work on Twitter conversations, describe steps to create a corpus for anaphora resolution, obtain predictions using Stanford statistical coreference system (Clark and Manning, 2015) and present analysis on the mediocre performance of the system on the Twitter corpus.

Entity resolution for email threads can also be cast as a multi-document or cross-document (CD) problem, where

[1] https://github.com/paragdakle/emailcoref

each email in an email thread represents a single document. EECB (Lee et al., 2012) and ECB+ (Cybulska and Vossen, 2014) are the most widely used corpora for this task setting. However, as seen with most corpora for within document entity resolution task, the documents in both corpus are taken from Google News search and hence are restricted to the news domain. Additionally, this corpus is primarily event-centric as it is based on the EventCorefBank (Bejan and Harabagiu, 2010). As per our knowledge, the model proposed by Barhom et al. (2019) is the current state-of-the-art on the ECB+ corpus. Barhom et al. (2019) model the event and entity coreference problem jointly and to show the differences between the tasks. This work also evaluates modeling coreference resolution for email conversations in a cross-document setting.

3. Task

We now formally define the entity resolution task for email threads. Let *T* be an email thread containing *N* email messages and *M* be the set of all mentions in *T* and *E* be the number of unique entities present in *T*. Let *C* be a set of chains of mentions $\{c_1, c_2, ..., c_E\}$, where each chain contains mentions referring to a unique entity. The term chain is analogous to a coreference cluster. Here, an entity belongs to one of the following classes: Person (PER), Organization (ORG), Location (LOC) or Digital (DIG). Section 4.3. further explains these classes. Compared to the CoNLL 2012 Shared Task, all singleton chains in *T* are considered to be a part of *C*. A singleton chain contains a single mention. Therefore, given an email thread *T*, the entity resolution task is to identify *C*.

4. Seed Corpus

4.1. Enron Email Corpus

The Enron Email Corpus[2] (Klimt and Yang, 2004) is one of the few publicly available email corpora containing actual user interactions. It was created by the CALO Project [3] (A Cognitive Assistant that Learns and Organizes). The corpus contains emails of 150 employees, organized in a directory structure. Each employee directory is further organized into folders like inbox, drafts, deleted_items, sent_items and other folder created by the employee. Over the years, contents of the corpus have been filtered to remove sensitive information like names, emails or attachments.

Some folders in each employee directory in the corpus contain auto-generated emails. For this research, only emails present in the "inbox" folder for each user have been considered. Table 1 shows the distribution of email threads from the "inbox" folder in terms of email messages.

4.2. Extraction and Filtering

An email thread is similar to a chat thread where multiple email messages have been exchanged over similar or extended topics. However, there can be cases where an email thread may end on an unrelated topic. For this corpus, email threads that satisfy the following constraints have been considered:

No. of Email Messages	Thread Count
1	30904
2-3	10015
4-7	2223
8-10	193
11-15	58
16-20	8
21-30	3
41	2

Table 1: Email thread count distribution in terms of email messages

- The thread must consist of more than three email messages. The objective of this research is to address both intra-email and inter-email entity resolution problems. Entity resolution in single email messages is also an unexplored problem, however, it is not being considered as a part of this research.
- More than half of the email messages in the thread must have some text body. The Enron Corpus contains multiple email threads having one email message being forwarded multiple times. Since such threads do not contain substantial text bodies, they can be discarded for this task.

To create an annotated seed corpus, a subset of 46 email threads containing 245 email messages is selected by first randomly choosing 16 users and then considering all the emails of those users satisfying the above constraints.

4.3. Annotation

For this annotation procedure, an *Entity* is defined as an object or a set of objects in the world. A *Mention* is defined as a span of text that refers to or mentions a real-world entity. For the scope of this task, the following entity types[4] are considered:

1. Person (PER): A single individual or a group of individuals can be annotated as a Person. A Person can be specified by name (John Doe), email address (johndoe@abc.com), first name (John), last name (Doe), occupation (the accountant), family relation (dad), pronoun (he), etc., or by a combination of these. All fictional human characters appearing in movies, TV, books, etc., are to be considered as a Person entity. A group of individuals that do not meet the requirements for an Organization entity, can be annotated as a Person entity. For example, "Analysts", "IBM's lawyers", "The family", "The house painters", etc.
2. Location (LOC): Places defined on a geographical basis and those that constitute a political entity are Location entities. An address, one-dimensional location like a border between two other locations, water-body, natural land-regions, non-named locations ("southern

[2] https://www.cs.cmu.edu/~./enron/

[3] http://www.ai.sri.com/project/CALO

[4] https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf

Africa") and general regions like "part of the city", "airspace", etc.

3. Organization (ORG): An organization entity must have some formally established association. Typical examples are businesses, government units, sports teams, and formally organized music groups. A department inside a company can also be termed as an organization.

4. Digital (DIG): A digital entity is a media or pointer to a media which is present on some form of digital storage. For example, email attachments, URLs, directory addresses.

When marking a mention, the following guidelines are observed:

- The part of speech of a mention can be one of *Nouns*, *Noun Phrases* and *Pronouns*.

- For the scope of this research, **no** event or verb is to be annotated.

- **No** date, time or date-time is to be annotated.

- When deciding on the width of a mention, the shortest width which describes the entity is chosen.

An email conversation, owing to the *To*, *Cc* and *Bcc* fields, can result in having participants with different levels of involvement. The participants in the *To* field are deemed to be directly involved in the conversation and those in *Cc* and *Bcc* to be indirectly involved. Pronouns such as 'you', 'team', 'everybody', and 'your' are considered to refer to each direct participant individually. This approach is similar to the one followed by Zhou and Choi (2018) in their work on resolution of plural mentions.

As compared to the OntoNotes 5.0 corpus, annotations for singleton chains are present in the seed corpus to help the model understand the email address and name mentions in the email header during fine-tuning. As an exception to this, singleton pronoun chains are excluded.

The annotation process for the seed corpus was carried out manually as a two-step process: identifying the mentions and chaining them. For the complete process, three annotators were used. Inter-annotator agreement on the Fleiss et al. (2003) Kappa statistic was $\kappa = 0.87$. A high κ value is due to a large number of email and name occurrences in the seed corpus which are unambiguous. All cases where no agreement was reached were resolved by discussion.

Table 2 gives details on the size of annotations in the seed corpus. The distribution of mentions per entity type is given in Table 3.

Note that the seed corpus also contains speaker annotations. Before manually annotating the seed corpus, an evaluation of weakly annotating email threads using the model proposed by Joshi et al. (2019b) with few manually annotated samples was carried out. Poor performance on this evaluation led to manually annotating the seed corpus.

Email Threads	46
Email Messages	245
Coreference Chains	866
Annotated Mentions	5834
Annotated Pronouns	981
Length of longest coreference chain	77
Average Length of coreference chains	6.73
Singleton chains	106

Table 2: Details on the size of annotations in the seed corpus

	PER	ORG	LOC	DIG
Mentions	76%	14%	4%	6%
Unique Entities	69%	17%	8%	6%

Table 3: Mention and entity distribution per entity type

5. Coreference Resolution in Email Conversations

The problem of anaphora resolution for Twitter conversations (Aktaş et al., 2018) exhibits characteristics similar to the problem in consideration. Email conversations are similar to Twitter conversations in terms of the tree-structure which is constructed by the 'reply-to' nature of the conversation. Furthermore, Twitter handles are analogous to email addresses and retweeting to forwarding. Lastly, both emails and tweets show some basic structure as to a header and body being present in every sample.

Yet, there are numerous differences between the two. Firstly, the use cases that the two mediums serve are very different. Twitter, is a microblogging and social networking platform, in which by default, all conversations are public and intended for a much larger audience. An email or "electronic mail" on the other hand is intended, like regular mail, directly for just the recipient individually or as part of a group. Secondly, the language in tweets uses many character reducing strategies or *textisms* (Lyddy et al., 2014) due to the character limit constraint. Since there is no explicit word limit set for a single email, the text is often more elaborate and descriptive.

A description of the challenges concerning coreference resolution observed in the seed corpus and general email conversations is provided below. Note that since this work deals with evaluating the problem of entity coreference resolution in email conversations, deriving solutions to these challenges has been left as future work.

5.1. Email addresses

An email address is a unique identifier for every user having an email account and hence can be considered as a mention representing a Person or an Organization entity. An email message in its entirety, that is header and body, most certainly contains the email addresses of the sender and recipient(s). However, it may or may not contain the names of the sender and/or recipient(s). Thus, it is crucial to identify and link an email address to the entity it represents.

Generally, email addresses bear some lexical similarity to the name of the entity it represents, but there are also instances when there is no overlap between the name of the entity and its email address. Additionally, an email address can represent a group of individuals as a whole. Such email addresses are called aliases. The difficulty of tracing conversations increases when an alias is involved in an email conversation. Example 1 shows various types of email addresses along with the corresponding name of the entity, if available, it represents.

Example 1. Examples of different types of email addresses along with the names of their corresponding entities

```
g..barkowsky@enron.com
Barkowsky, Gloria G.

theresa.staab@enron.com
Staab, Theresa

smu-betas@yahoogroups.com
SMU Betas

fackel@yahoo.com
Leah
```

5.2. Different email thread structures

A consistent email header, body and thread structure eases the pre-processing task and extraction of various features from emails. It also helps in faster error analysis. The emails in the Enron corpus have varied header as well as email thread structures. Example 2 shows a few examples of the different email headers seen in the Enron corpus.

Example 2. Few examples of different email headers present in the Enron Corpus

```
1.
Message-ID: <16657248.1075852679695..
Date:  Fri, 17 Aug 2001 11:40:43 -0700
(PDT)
From:  chuck.paul@a-closer-look.com
To:  smu-betas@yahoogroups.com
Subject:  [smu-betas] FW: ...
Mime-Version:  1.0
Content-Type:  text/plain..
Content-Transfer-Encoding:  7bit
X-From:  "Chuck Paul"
<chuck.paul@a-closer-look.com>
X-To:  SMU/Beta's <smu-betas..
X-cc:
X-bcc:
X-Folder:  JSKILLIN (Non-Privileged)..
X-Origin:  Dasovich-J
X-FileName:  JDASOVIC..

2.
From:  Miller, John
[mailto:miller@advlaser.com]
Sent:  Friday, August 17, 2001 2:33 PM
To:  'chuck.paul@a-closer-look.com'
Subject:  RE: A friend thinks you ..

3.
----- Forwarded by Sheila
Rappazzo/OGW/NYSDPS on 10/05/01 01:11
PM -----

4.
"Melissa L. Lauderdale"
<lauderdale@bh-law.com> 10/05/01 12:27
PM
To:  "'sheila_rappazzo@dps.state.ny.us'"
<sheila_rappazzo@dps.state.ny.us>,..
cc:
Subject:  RE: Oct 10th Meeting ..
```

A fixed structure of an email thread plays an important role in deciding email boundaries and thereby the scope of different pronouns that are local to an email message in the thread. Threads in the Enron Corpus generally follow a time based last to first ordering. However, multiple instances of out of order threads as well as different email structures are seen. Example 3 shows one such structure.

Example 3. An example of an out of order email thread structure present in the Enron Corpus

```
...
-----Original Message-----
Sent:  Friday, May 25, 2001 12:35 PM
...
email contents
....
---------------------- Forwarded
by Jaime Sanabria/ENRON_DEVELOPMENT
on 05/25/2001 12:42 PM
---------------------------
on 05/21/2001 03:49:00 PM
To:  "ENRON: Sanabria, Jaime"
<jaime.sanabria@enron.com>
...
```

5.3. Name abbreviations and variations

Identifying the name of a PER or an ORG entity is crucial not only for correct coreference chain identification but also for tasks like anaphora resolution or question answering. The semi-structured nature of email messages adds to the complexity of identification of all names referring to the same entity. The names present in the email message headers for PER type are either full names or the names that are registered in the system. However, in an email message body, name abbreviations or variations are used between frequent or known participants. For an ORG entity, the signature found at the end of an email message contains a non-abbreviated version of the name as compared to the names found in the message subject or body. Examples 4 and 5 shows a few name abbreviations and variations observed in the seed corpus.

Example 4. *Frazier, Perry* referred to as *PT*, *Kimberly* as *Kim*, *Miller, Mary Kay* as *MK* and *Transwestern Commercial Group* as *TW.*

Example 5. *Robert Superty* ⟷ *Bob Superty* and *William E. Brown* ⟷ *Bill Brown*.

5.4. General challenges

Some of the general challenges involved in working with email conversations are:

1. Typos affecting referring expressions.

   Example 6. *They will also be proposing that the Commission switch from long run marginal cost to embedded cost principles for allocating costs of service among its customers.* ***The*** *also propose a $187 million or 12.5% rate increase annually, compared to present rates.*

2. Speaker references: Email conversations are multi-user conversations by nature. Due to this, third-person pronouns are used very frequently. Aktaş et al. (2018) is the only work in our knowledge considering this phenomenon in a conversational setting. Although an email thread can be viewed as a turn-based sequential conversation over time, the time sequencing may not align with the flow of the conversation, thereby adding to the complexity of the task.

3. Ambiguity with first-person plural pronouns: In an email conversation, especially in a formal setting, the participants represent a larger group or an organization. These cases add ambiguity to the resolution of first-person plural pronouns. Consider the pronoun '*we*', it can resolve to both the sender and recipient together or the entity the sender is representing.

6. Experiments

6.1. Models

Entity resolution for the seed corpus is evaluated by considering both within document (WD) and cross-document (CD) formulations of the task. For the WD formulation, the model proposed by Joshi et al. (2019b) is used. This is the current state-of-the-art for the CoNLL 2012 shared task. Joshi et al. (2019b) take the c2f-coref model (Lee et al., 2018) as the base model. The proposed model uses BERT to obtain the span embedding replacing the original LSTM-encoder in the c2f-coref model. For each mention embedding, a mention score is computed and for each valid antecedent:mention pair, an antecedent score is computed. Eventually, using these scores, the probability of an antecedent belonging to a chain is computed. The model was fine-tuned and evaluated on the CoNLL 2012 shared task and GAP (Webster et al., 2018) corpus. As compared to using simple BERT as a model component, Joshi et al. (2019a) show that replacing BERT with SpanBERT leads to better performance on the CoNLL task. For readability, this model has been termed as **OntoSpanBERT** and the SpanBERT model not fine-tuned on the OntoNotes 5.0 corpus as **VanillaSpanBERT** respectively.

For the CD formulation of the task, the model proposed by Barhom et al. (2019) is used. The model was trained jointly on ECB+ corpus for both event and entity resolution tasks and is the current state-of-the-art for the ECB+ corpus. The model iteratively performs event and entity coreference resolution. The results of each subtask are alternately used to merge predicted chains in each iteration. The authors use mention lexical span, surrounding context, and event-entity mention relations via predicate-arguments structures to obtain predictions.

6.2. Setting

For the corpus evaluation in the WD setting, the independent variant of OntoSpanBERT [5] has been used. Since, the original CoNLL 2012 task does not include singleton chains in its training and predictions, during computation of performance metrics, the scores with and without singleton chains in the corpus are reported. The input to the models is in CoNLL 2012 format with appropriate pre-processing done for each model. The experiments were run on a GPU environment comprising of 8 cores of Nvidia GTX 1080 Ti with 12 GB of memory per core.

6.3. Evaluation

The majority of the recent work done on entity coreference use the MUC, B^3 and CEAFE metrics (Pradhan et al., 2012). Moosavi and Strube (2016) show the shortcomings of each of these metrics and propose the Link-based Entity Aware (LEA) metric[6]. The results using all four metrics, for comparability as well as correctness, have been reported. The official scorer [7] provided by CoNLL 2012 shared task is used. The official scorer raises a non-crashing duplicate reference error when a single-token mention belongs to more than one chain. This error is also observed on the OntoNotes corpus and hence this paper reports the scores ignoring the errors.

Both models deal with more entity types than those defined here like Facility, Event, Product or Vehicle. However, since the model does not output the entity type of a chain, no chain from the predictions is removed. Furthermore, since none of the models were trained on the DIG entity type, scores excluding annotations for that type have also been reported. Table 4 shows the empirical results of the experiments. For the WD setting of the task, the OntoSpanBERT performs best when the test data contains only PER, ORG and LOC entity types and no singleton chains. The +0.99 F1 increase after removing singleton chains is understandable as removing singleton chains leads to a reduction in the size of expected final chaining resulting in a higher recall as compared to the general setting. Likewise, the drop in precision for all metrics after removing the DIG entity type shows that the current model already captures the type even if the training corpus did not contain annotations for the DIG type.

[5] `https://github.com/mandarjoshi90/coref`

[6] Zhou and Choi (2018) propose variations of B^3 and CEAFE which may be more appropriate here since this work follows a similar annotation scheme. However, for ease of comparison, we skip using these variations.

[7] `https://github.com/conll/reference-coreference-scorers/tree/LEA-scorer`

Next, a VanillaSpanBERT is fine-tuned on the annotated seed corpus using an 80:20 train-test split. This model is referred to as **SeedSpanBERT**. The results obtained show that SeedSpanBERT exhibits the best performance. However, it is important to note that the test set contains merely 10 email threads. Distant supervision (Mintz et al., 2009) can be used to create a larger corpus for the task but is left as future work.

Note that fine-tuning OntoSpanBERT on the seed corpus was attempted and it did not result in any improvement in the results that were observed before the additional fine-tuning. The small size of the annotated seed corpus results in only slight weight perturbations, which is not significant enough to change the predicted chaining.

In a cross-document formulation, an email thread is viewed as a collection of email messages. Since the seed corpus does not include event annotations, the predictions obtained were not meaningful and thus, could not be used for evaluation.

7. Error Analysis

Error analysis presented here has been performed on the predictions of OntoSpanBERT and SeedSpanBERT. The predictions obtained by the OntoSpanBERT model were assessed with different variations of the seed corpus (without singletons, DIG entity type or both). There are five general types of errors observed. Primary error analysis is performed on the predictions of the OntoSpanBERT model and changes observed in the predictions on the test set using SeedSpanBERT are reported. Table 5 gives more information on the statistics of each error type. It is important to note that since the error categories are not mutually exclusive, the possibility of a span of text contributing to more than one error category exists. The objective here is to get an insight into the type of errors observed and the individual statistics.

Figure 1 shows a comparison between OntoSpanBERT and SeedSpanBERT in the distribution of the first and fifth category errors per entity type on the test set. The comparison considers only email threads in the test set. It can be seen that a high percentage of PER mentions in the corpus largely influence the learning of the model. Additionally, since the seed corpus compared to the OntoNotes 5.0 corpus does not contain sufficient ORG mentions, OntoSpanBERT will likely perform better on ORG mentions. From the performance of the models on the DIG entity type, it can be inferred that OntoSpanBERT, in terms of mention span identification, does a better job at implicitly capturing the entity type.

7.1. Missing references in the chain

1. Missing references in the email header

   The English language section of the OntoNotes 5.0 corpus consists of texts from one of the following categories: newswire, magazine articles, broadcast news, broadcast conversations, web data, conversational speech data and English translation of the New Testament. None of these categories have texts with a header close or similar to an email message header. Additionally, the corpus contains no email addresses, thereby making an email address an unfamiliar concept to the trained model. 45% of missing references in the chain belong to this sub-category of which missing email address mentions are 56% and missing name mentions are 44%.

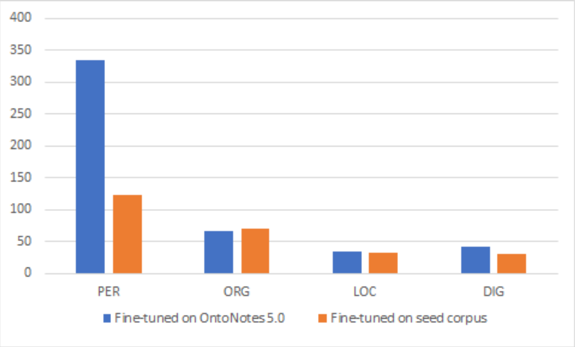

Figure 1: Comparison of error distribution per entity type between OntoSpanBERT and SeedSpanBERT

2. Missing pronomial references.

   This error sub-category contributes to 11% of this error category. Table 6 gives a distribution of the errors per pronoun category.

3. Other missing references in the chain.

   The final sub-category consists of the remaining missing references. These errors are further divided into their corresponding entity types to get a better understanding of the missing references. Table 7 shows the corresponding breakdown. The results show that even though the model was never trained on the DIG entity type, it partially or completely predicted 210 mentions out of 348 present in the seed corpus.

   From the results of SeedSpanBERT, it can be inferred that the training process helped the model learn the importance of email headers as well as focus on the relevant entity mention types. It does not help largely in solving the mention identification problem in the rest of the email body.

7.2. Decomposition of a single chain

This error category represents the cases when a single coreference chain in the annotated corpus was present as multiple chains in the predictions. However, taking a union of all these chains does not necessarily result in obtaining the original annotated chain. Of the chains which are decomposed, 58% are split into two parts, 33% into three parts, and 8% into four parts. One instance of decomposition into five parts is seen. Even though there is a reduction in the number of decomposed chains with SeedSpanBERT, the fine-tuning process results in creating longer chains composed of multiple single entity chains. Singleton chains are the dominant ones to be absorbed in other chains. The small size of the seed corpus is a factor that attributes to this behavior.

	MUC			B^3			CEAFE			LEA			
Model	P	R	F1	P	R	F1	P	R	F1	P	R	F1	Avg. F1
OntoSpanBERT	58.7	40.8	46.5	50.9	23.1	30.1	29.7	30.7	28	47.6	20.6	27	34.9
w/o singletons	58.7	40.9	46.6	50.8	23.6	30.5	29.6	34.5	29.8	46.2	20.4	26.8	35.6
w/o DIG	57.6	42.2	47.2	49.7	23.8	30.4	28.9	31.4	27.8	46.2	21	27.3	35.2
w/o both	57.6	42.2	47.2	49.6	24.2	30.9	28.7	**35.2**	29.4	48.3	19.6	26.2	35.9
SeedSpanBERT	**79.5**	**64.8**	**70.9**	**62.3**	**46.6**	**52.1**	**57.4**	31.7	**39.1**	**57.4**	**42.6**	**47.8**	**54**

Table 4: Evaluation results for OntoSpanBERT and SeedSpanBERT on the seed corpus. Avg. F1 score is computed using MUC, B^3 and CEAFE metrics

Category	Count$_f$	Count$_1$	Count$_2$
Missing references in the chain	1587	476	254
Decomposition of a single chain	135	42	22
Wrong items in the chain	245	90	254
Missing chains	222	65	27
Incorrect and irrelevant mention spans	1054	271	27

Table 5: Statistical information of errors observed. Count$_f$ column reports numbers observed on the full seed set, Count$_1$ on the test set with OntoSpanBERT, and Count$_2$ on the test set with SeedSpanBERT respectively

	1st Person	2nd Person	3rd Person	Other
%	36%	52%	2%	10%

Table 6: Distribution of missing references per pronoun type

	PER	ORG	LOC	DIG
%	37%	31%	13%	19%

Table 7: Distribution of missing references per entity type

Type	% change
Missing references in the chain	-46
Missing references in the email header	-78
Missing email references	-77
Missing name references	-80
Missing pronomial references	-63
Other missing pronomial references	-19
Decomposition of a single chain	-48
Wrong mentions in the chain	+182
Pronouns	$+2220_1$
Other PER entity mentions	$+475_2$
Missing chains	-59
Chains of length 2	-86
Incorrect and non-relevant mention spans	-92
Incorrectly identified mention spans	-65
Non-relevant mention spans	-93
Duplicate name mention spans	-100

Table 8: Detailed statistics of error reductions with SeedSpanBERT as compared to OntoSpanBERT. 1: Count increased from 5 to 116. 2: Count increased from 4 to 23.

7.3. Wrong mentions in the chain

This error category indicates that an incorrect mention is identified as part of a coreference chain. On the entire seed corpus, the majority of the errors are pronouns (68%) and other PER entity mentions (17%). Post-fine-tuning the number of wrong mentions in the test set increase by 182%. Merging chains of length 2-3 into bigger chains or other chains of similar lengths is the primary factor for this significant increase. The scores on the MUC metric show that the OntoSpanBERT model does a better job at chaining mentions but lacks the knowledge of identifying mentions in an email corpus. On the other hand, the SeedSpanBERT model, post-fine-tuning on the seed corpus, learns how to identify mentions but fails at the chaining task.

7.4. Missing chains

Since the CoNLL 2012 shared task corpus does not contain singleton chains, they have been excluded from this error category. Additionally, chains that have only one of their elements predicted are tagged as missing chains. Table 9 shows the breakdown of the missing chains by the length of the individual chains respectively. Chains of length between 2 and 3 dominate this error sub-category. Most of these chains consist of an email address and the corresponding name of the entity referred only in the header of one email message in the email thread. Few examples of these types of chains are *['dutch.quigley@enron.com', 'Quigley, Dutch'], ['ed.mcmichael@enron.com', 'McMichael Jr., Ed']*. The results of SeedSpanBERT do not imply that the entire chain is present as expected in the predictions, but instead implicates that the elements of a previously missing chain are present either in a single chain or as parts of another chain.

Length	2-3	4-5	6-10	10+
Count	172	27	17	6

Table 9: Breakdown of missing chains by the length of the chain

7.5. Incorrect and irrelevant mention spans

1. Incorrectly identified mention spans

These types of errors consist of predicted mention spans whose width differs from the expected mention spans. Email headers consist of full names of the sender as well as recipients. However, the full span of these names is not predicted on multiple occasions (71%). Example 7 consists of a sample prediction where *Barkowsky, Gloria G.* was the expected mention prediction but the system returned *Gloria G.*.

Example 7. Example showing partial and duplicate name mention predictions

```
Date:  Mon, 17 Dec 2001 14:28:03
-0800 (PST)
From:  g..barkowsky@enron.com
To:  theresa.staab@enron.com
Subject:  RE: Final Statements and
Invoices for November
X-From:  Barkowsky, Gloria G.
X-To:  Staab, Theresa
X-cc:
X-bcc:
yes, I'll do this.
Do you have anything for Crestone
and Lost Creek?
```

2. Irrelevant mention spans

   The SpanBERT model used for obtaining predictions is fine-tuned on the CoNLL 2012 shared task which consists of additional entity types. This results in many additional or irrelevant mentions being predicted by the model, that are considered as errors. One of the contributing factors to the increase in precision of SeedSpanBERT was the 93% reduction seen in these spans. Fine-tuning the model on the seed corpus helped in excluding the learning of the other entity types present in OntoNotes 5.0 corpus, thereby not predicting spans representing those types.

3. Duplicate name mention spans

   When a sub-span of a name mention span is predicted as part of another chain or the same chain, the sub-span is considered to be a duplicate one as the full name is considered to be the entity representative span. Example 7 shows the scenario where the expected mention is just *Staab, Theresa*, but *Theresa* is also predicted as a mention span.

8. Conclusion

Entity coreference resolution in email threads is an unexplored problem. The paper contends that the problem is an important one and elaborates on the uniqueness as well as challenges of the same. The construction of the seed annotated corpus and reported statistics are explained, highlighting the complexity of the seed corpus. The problem is evaluated in a within document setting. Predictions obtained on the seed annotated corpus show that the current state-of-the-art models exhibit a less than average performance. Also, the performance of fine-tuning on the seed corpus is reported. Qualitative error analysis on the predictions of the SpanBERT model (Joshi et al., 2019b) and conclusions drawn from the analysis are presented. In the future, the construction of a larger corpus using the seed corpus will be explored for identifying solution(s) for the entity resolution problem in email threads. This will be followed by a performance evaluation of the new solution(s) on the existing corpora.

9. References

Aktaş, B., Scheffler, T., and Stede, M. (2018). Anaphora resolution for twitter conversations: An exploratory study. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 1–10.

Barhom, S., Shwartz, V., Eirew, A., Bugert, M., Reimers, N., and Dagan, I. (2019). Revisiting joint modeling of cross-document entity and event coreference resolution. *arXiv preprint arXiv:1906.01753*.

Bejan, C. and Harabagiu, S. (2010). Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1412–1422.

Chen, Y.-H. and Choi, J. D. (2016). Character identification on multiparty conversation: Identifying mentions of characters in tv shows. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 90–100.

Chinchor, N. A. (1998). Overview of muc-7/met-2. Technical report, SCIENCE APPLICATIONS INTERNATIONAL CORP SAN DIEGO CA.

Clark, K. and Manning, C. D. (2015). Entity-centric coreference resolution with model stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1405–1415.

Cohen, W. W., Carvalho, V. R., and Mitchell, T. M. (2004). Learning to classify email into âspeech actsâ. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 309–316.

Cybulska, A. and Vossen, P. (2014). Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *LREC*, pages 4545–4552.

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S. M., and Weischedel, R. M. (2004). The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, page 1. Lisbon.

Fleiss, J., Levin, B., and Paik, M. (2003). The measurement of interrater agreement. In *Statistical Methods for Rates and Proportions, Third Edition*, pages 598 – 626. John Wiley Sons, Inc.

Grishman, R. and Sundheim, B. (1996). Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Guillou, L. (2012). Improving pronoun translation for statistical machine translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the*

European Chapter of the Association for Computational Linguistics, pages 1–10. Association for Computational Linguistics.

Hendrickx, I. and Hoste, V. (2009). Coreference resolution on blogs and commented news. In *Discourse Anaphora and Anaphor Resolution Colloquium*, pages 43–53. Springer.

Ide, N., Baker, C., Fellbaum, C., Fillmore, C., and Passonneau, R. (2008). Masc: The manually annotated subcorpus of american english. In *6th International Conference on Language Resources and Evaluation, LREC 2008*, pages 2455–2460. European Language Resources Association (ELRA).

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2019a). Spanbert: Improving pre-training by representing and predicting spans. *arXiv preprint arXiv:1907.10529*.

Joshi, M., Levy, O., Weld, D. S., and Zettlemoyer, L. (2019b). Bert for coreference resolution: Baselines and analysis. *arXiv preprint arXiv:1908.09091*.

Klimt, B. and Yang, Y. (2004). The enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning*, pages 217–226. Springer.

Lee, H., Recasens, M., Chang, A., Surdeanu, M., and Jurafsky, D. (2012). Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 489–500. Association for Computational Linguistics.

Lee, K., He, L., and Zettlemoyer, L. (2018). Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana, June. Association for Computational Linguistics.

Lyddy, F., Farina, F., Hanney, J., Farrell, L., and Kelly O'Neill, N. (2014). An analysis of language in university students' text messages. *Journal of Computer-Mediated Communication*, 19(3):546–561.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

Moosavi, N. S. and Strube, M. (2016). Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., and Zhang, Y. (2012). Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40. Association for Computational Linguistics.

Recasens, M., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., Poesio, M., and Versley, Y. (2010). Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8. Association for Computational Linguistics.

Soboroff, I., de Vries, A. P., and Craswell, N. (2006). Overview of the trec 2006 enterprise track. In *Trec*, volume 6, pages 1–20.

Ulrich, J., Murray, G., and Carenini, G. (2008). A publicly available annotated corpus for supervised email summarization. In *Proc. of aaai email-2008 workshop, chicago, usa*.

Webster, K., Recasens, M., Axelrod, V., and Baldridge, J. (2018). Mind the gap: A balanced corpus of gendered ambiguou. In *Transactions of the ACL*, page to appear.

Zhou, E. and Choi, J. D. (2018). They exist! introducing plural mentions to coreference resolution and entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 24–34.