The Challenges of Cross-Document Coreference Resolution in Email

Xue Li
x.li3@uva.nl
University of Amsterdam
Amsterdam, Netherlands

Sara Magliacane
s.magliacane@uva.nl
University of Amsterdam
MIT-IBM Watson AI Lab
Amsterdam, Netherlands

Paul Groth
p.t.groth@uva.nl
University of Amsterdam
Amsterdam, Netherlands

ABSTRACT

Long-form conversations such as email are an important source of information for knowledge capture. For tasks such as knowledge graph construction, conversational search, and entity linking, being able to resolve entities from across documents is important. Building on recent work on within document coreference resolution for email, we study for the first time a cross-document formulation of the problem. Our results show that the current state-of-the-art deep learning models for general cross-document coreference resolution are insufficient for email conversations. Our experiments show that the general task is challenging and, importantly for knowledge intensive tasks, coreference resolution models that only treat entity mentions perform worse. Based on these results, we outline the work needed to address this challenging task.

CCS CONCEPTS

• **Computing methodologies → Information extraction**.

KEYWORDS

cross-document coreference resolution, email conversations, entity resolution, challenges, conversational data

ACM Reference Format:
Xue Li, Sara Magliacane, and Paul Groth. 2021. The Challenges of Cross-Document Coreference Resolution in Email. In *Proceedings of the 11th Knowledge Capture Conference (K-CAP '21), December 2–3, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3460210.3493573

1 INTRODUCTION

Coreference resolution (CR), the task of determining which textual mentions refers to the same entity, is a long standing and important task in natural language understanding (NLU) [8]. Being able to perform coreference resolution is particularly important in knowledge capture settings dealing with emergent entities or entities which are not represented in existing knowledge graphs [13, 16].

One such setting is knowledge capture from long form conversations such as medical conversations [18], personal dialogues [14], threaded discussion forums [20], and email conversations [4]. CR in conversational data is challenging because of the change in speakers and high lexical ambiguity [20]. Such lexical ambiguity is caused by the often large amounts of shared background knowledge needed to understand the conversation [17]. For example, imagine an email thread discussing the edits to a document. The participants know the contents of the document and its intended purpose, but such knowledge is not easily available to an extraction system.

In this work, we study CR performance on a type of long form conversation, namely, email conversations. Email provides a challenging but realistic case because it contains more than one speaker, and has no limitation on the length of the emails. Surprisingly, there has not been a lot of work on CR in emails, and only recent work has begun to study CR on emails using deep learning models [3] based on a small hand-annotated subset of the Enron email corpus [9]. The corpus only contains coreference annotations for mentions within email and across emails in each thread. Dakle et al. [3] formulate this problem as Within-Document Coreference Resolution (WD-CR), treating an email thread as a single document, and ran a WD-CR state of the art method, spanBert [6]. Given that entities are often referred to across emails and between multiple threads, we instead consider a cross-document formulation more natural. Therefore, we evaluated a state of the art cross-document coreference resolution (CD-CR) [1] model on the same corpus, considering each email in a thread as a single document.

Our results show that email conversations are a particularly challenging domain, where the state-of-the-art performs significantly worse than when applied to commonly-used corpora in the field such as news (e.g. CoNLL F1 34.4 on news vs. 27.4 on email). Based on these results, we outline the key challenges for CR in emails and paths forward to addressing them. In summary, the contributions of the paper are as follows:

(1) performance results for the state-of-the-art CD-CR on email conversations including an ablation study investigating performance with and without pronominal coreference;
(2) a qualitative error analysis that identifies where the challenges in this domain arise from; and
(3) paths forward to addressing this important domain.

2 BACKGROUND

We briefly introduce the state of the art on coreference resolution (CR), focusing on the cross-document setting and email. We refer the reader to [5, 17] for recent reviews of approaches to CR.

State-of-the-art CR: In the early literature, CR systems typically contained two separate stages: mention extraction with existing feature-engineered systems (e.g. syntactic parsing) and coreference relation finding. The first work that joined the two stages was e2e-coref [10]. It proposed an end-to-end deep learning architecture that jointly learned to extract mentions and rank them as

K-CAP '21, December 2–3, 2021, Virtual Event, USA
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8457-5/21/12.
https://doi.org/10.1145/3460210.3493573

to whether they corefer. Later, c2f-coref [11] proposed a span refinement technique to iteratively refine the span representations of mentions to help address global inconsistency and achieve higher-order inference. The current state of the art in within-document CR, spanBert [6], leverages large pre-trained language models [7].

However, the success of such models relies on textual orders. In the cross-document (CD) setting, generally there is no temporal ordering between the documents, which impacts negatively the performance of these models [1]. Instead of learning an antecedent distribution and chaining the current mention to the most probable antecedent, the current state-of-the-art [1], which we evaluate in this paper, learns to extract candidate mentions and compare all the most probable mentions for potential coreference. We explain the model in more detail in Section 3.

To evaluate the performance of CD-CR, ECB+ (EventsCorefBank+) [2] is the most commonly used dataset in recent years. It consists of news articles organized by topics, and contains both WD and CD CR annotations for both entities and *events*. To address the lexical ambiguity challenge in CD-CR in particular, within each topic, each instance (also called *subtopic*) is a pair of similar but different events. ECB+ is a challenging dataset that focuses mostly on the specific situation for disambiguation for events that contain similar entities. In the rest of the paper, we show that CD-CR models that perform reasonably on ECB+ do not work well on email.

More broadly, Cattan et al. [1] highlighted the pressing need for more realistic evaluation on CD-CR task. In particular, existing work often reports performance using golden mentions instead of predicted mentions. Their work showed that the contribution of mention prediction is significant and should be considered in evaluation. Thus, an end-to-end training approach is more realistic for real-world data.

CR for email: Email conversations have long been studied on tasks such as classification, search and summarization [3]. One of the largest email corpus is the Enron Email Corpus [9] containing emails of 150 employees of the Enron Corporation. Surprisingly, only recently has CR within email come to the fore. In particular, [3] provided, what appears to be, the first analysis of entity coreference for email in the literature. They introduced a manually annotated seed corpus (SC) containing 46 threads and 245 emails from the Enron Email Corpus. The authors filtered the original corpus so that each thread contains more than 3 emails in order to have both within-email annotations as well as cross-email annotations, and each email has meaningful text body instead of forwarding messages. In their analysis they evaluated a state of the art model, spanBert [6], that they had fine-tuned to the seed corpus, which had 54 F1 score. This is 26 points below the state-of-the-art in general WD-CR [19].

Building of this work, Dakle and Moldovan [4] introduced a larger dataset called CEREC with 6001 email threads containing 36,448 emails with weakly labeled data. The labeling has two stages, mention identification annotations and coreference relation annotations. Both stages use pre-trained spanBert [6] to annotate without further training. The mention annotations produced by spanBert are then manually corrected. For the coreference relations annotation, first a small subset of email threads are manually annotated as a validation set for training performance. Then spanBert is trained on the manually annotated seed corpus [3] and then the coreference relations are obtained on the large dataset based on the golden mentions. Surprisingly, training a state-of-the-art WD-CR model on CEREC showed roughly no change in performance, with a reported F1 score of 54.1 [4] compared to the F1 score of 54 [3] on the seed corpus.

3 METHOD

We now describe the model and model training we use to characterize the performance of CD-CR in email.

Model architecture: As previously discussed, we use the state-of-the-art model architecture [1] and summarize it here:

The model contains three main modules: a `span_scorer`, a `span_embedder`, and a `pairwise_scorer`.

The model takes a set of documents as input and uses a pre-trained language model to get a contextual representation of each token in the document. It then segments the tokens to determine possible mention candidates up to a pre-defined mention width. Then, the `span_embedder` is used to obtain the embedding for each mention candidate. The model then prunes all candidates according to `span_scorer`. The most probable candidates are paired together and scored by the `pairwise_scorer`, which is a multi-layer perceptron (MLP). During inference, agglomerative clustering is used to return final clusters of coreferent mentions.

Cattan et al. [1] proposed three different training styles: pipeline, continue and end-to-end (e2e). The main difference is that they freeze different parts of the pipeline. We use the e2e style, which trains all three modules end-to-end. A binary cross-entropy loss is used to jointly train the `span_scorer` and `pairwise_scorer`.

Model training: We train this model on the SC corpus [3] introduced previously. We use the e2e training style to achieve a more realistic evaluation. For comparison to ECB+ results, we assume that each email thread is equivalent to an ECB+ topic. We split long documents to the maximum document length of $n = 512$ tokens.

We use a span representation based on e2e-coref [10]. First, we use pre-trained RoBERTa [12] to encode each token in the input. For each mention candidate, or *span*, i we compute the embedding s_{emb} as a concatenation of 4 different components:

$$s_{emb}(i) = (x_{start(i)}, x_{end(i)}, \hat{x}_i, \phi(i))$$

where $x_{start(i)}$ and $x_{end(i)}$ are the token representation of the first and last token in m_i,while $\hat{x}_i$ is the weighted sum of all token representations in span i (i.e. the attention), and $\phi(i)$ is the feature vector that encodes the length information of span i.

Then all encoded mentions will be scored by the mention scorer $s_m(\cdot)$, which are then pruned to retain only $\lambda = 35\%$ percent of mentions. We use a multiple layer perceptron (MLP) layer with ReLU activation function as our $s_m(\cdot)$. Then the pruned mention candidates will be paired up and scored by a pairwise scorer $s_p(i, j)$, where i, j might or might not from different documents. The pairwise scorer $s_p(i, j)$ is also a MLP. We perform negative sampling.

All three modules mention scorer $s_m(\cdot)$, span embedder s_{emb} and pairwise scorer $s_p(i, j)$ are jointly trained by optimizing the binary cross-entropy loss over pairs:

$$L = -\frac{1}{N}\sum_{i,j\in N} y \cdot log(s(i,j))$$

$$s(i,j) = s_m(i) + s_m(j) + s_p(i,j)$$

where N is the set of mention pairs and y indicates the binary label. When $y = 1$, it indicates the mention pair is coreferent.

Experimental settings The experiments are carried out on titan RTX 24GB GPU, it takes approximately 70 minutes to train and evaluate 10 epochs. Our training data is split into training, validation and test set in an 80:10:10 ratio. The seed corpus (SC) contains 43 email threads and 228 emails in total. All threads have at least 4 emails. The whole dataset contains 3815 mentions across emails within threads. Each email contains header, body and footer.

We note that emails have conversational features, i.e. the speaker of each email changes in a thread and therefore it is more confusing to learn the antecedent distributions for pronominal mentions across emails. Simple rules for *{I, you}* are easy to resolve, but for second order pronouns (e.g. our team) and above it is challenging to learn the distribution.

To further study cross-documents pronominal resolution, we first remove first order pronouns in a union of {I, you}, and then remove a whole list of pronouns [1] from the dataset. Table 1, details the number of removed mentions.

Mention type	Number	Subtracted
All mentions	3815	0
- subset of pronouns {i, you}	3298	517
- all pronouns	2613	1202

Table 1: Number of mentions before & after removing pronouns.

4 RESULTS

We evaluate the model on the standard coreference resolution metrics, including MUC, B^3, CEAFe, LEA. The results are reported in Table 2. We compare our cross-document coreference resolution model on the test set of the email seed corpus with the ECB+ test set. To make the results comparable, we compare the evaluation result, on the ECB+ test set for entity resolution only, with predicted mentions, and on a topic level. The final F1 score is 34.4. For the email seed corpus, we first evaluate our model on the full test set with all pronouns. The F1 score is around 27.4, which is a significant 7 point drop. Then, a subset of pronouns $\{I, you\}$ are removed from the mentions. This has a slightly worse F1 score of 26.7. After the removal of all pronouns, the model produce an F1 score of 23.5.

This shows that the model is able to learn some alignment between pronouns but performs worse on less generic mentions, which often characterize entities.

5 CHALLENGES

We now discuss the main challenges faced in order to improve coreference resolution in email using a qualitative error analysis. Subsequently, we articulate several directions for future work.

Informal language: As mentioned in Section 2, most datasets for the CD-CR task are in the *news* domain, where the language style is more formal and structured. Hence, most entity mentions have previous references. In comparison, email is more informal and hence less structured. Table 3 shows examples of this from the ECB+ and Email datasets. In the ECB+ example, we can see that the text is clear. The main coreference challenge is that two different events have similar names. In addition, in the Email example, the coreference relations are more complicated. In the Document 1, one would need header information to reason what *we* refers to. Similarly, the two emails are needed jointly to understand the coreference for *your* in Document 2 which must exclude the speaker herself. This example illustrates that treating coreference resolution in emails as within document is insufficient.

Variety of surface forms: Table 1 shows pronoun-related mentions take up to 31.5% of all mentions. As discussed in Section 4, model performance drops by 4 points on the F1 score after removing all pronouns. This indicates that the model struggles on predicting coreference between other type of mentions. Some example mentions are shown in Example 1.

Example 1

```
'Frazier,Perry'
'Perry'
'FP'
'perry.frazier@enron.com'
```

From the example, we can see that there are multiple different surface forms for the same entity and these forms vary widely. Models need to become better at coping with this sort of wide variation. This is a known challenge in the literature [17] but given the nature of email appears with more frequency.

Sparsity: The prior two challenges are exasperated by the fact that there is a lack of data necessary to train good models. Unfortunately, the current weakly labeled CEREC dataset is inadequate due to low quality annotations.

6 PATHS FORWARD

More data: A clear path forward is the provision of more annotated data. Here, we suggest that instead of using an existing coreference model to generate data as in CEREC, a data programming approach [15] might be more appropriate.

Incorporation of rules: While, as we have discussed, email conversations are complex, there are opportunities to take advantage of common patterns within conversations. For example, we calculated that using simple rules to align subset of pronouns $\{I, you\}$ with email headers could resolve around 13.6% of mentions. Incorporating these and other rules with current models is a promising direction.

Language models for email text: Creating a pre-trained model specifically for emails could help to better capture the unique idiosyncrasies of email. This could also address the huge amount of memory needed for token encoding.

Better pruning strategies: The current state-of-the-art model on CD-CR, currently pairs up all mention candidates. This creates a massive search space thus requires a pruning factor to be given beforehand. Dynamically pruning candidates with a smarter strategy could reduce this space of potential candidates.

[1] The full list of pronouns that we removed is here: https://gist.github.com/effyli/da7c4243f296a6c689697384b48896f5

	MUC			B^3			CEAFe			LEA			CoNLL
	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1	F_1
ECB+ Entities Test set	41.7	52.3	46.4	24.8	37.1	29.7	27.4	26.8	27.1	22.3	34.4	27.1	**34.4**
Email Test data	41.2	25.0	31.1	40.4	22.5	28.9	37.1	15.9	22.3	26.3	12.0	16.5	**27.4**
- Subset of pronouns $\{I, you\}$	41.7	24.7	30.9	40.1	23.7	29.9	31.3	14.0	19.3	20.6	11.7	14.9	26.7
- All pronouns	34.8	30.0	32.2	21.0	37.6	27.0	28.5	6.8	11.0	10.0	18.0	12.9	23.5

Table 2: Cross-document coreference results. ECB+ Test Set with entities only on the topic level as a baseline (from [1]). Email test set [3], with/without(-) subset($\{I, you\}$) or removing all pronouns.

Dataset	Doc 1	Doc 2
ECB+	News that Barack Obama may name Dr. Sanjay Gupta of Emory University and CNN as his Surgeon General has caused a spasm of celebrity reporting.	President Obama will name Dr. Regina Benjamin as U.S. Surgeon General in a Rose Garden announce ment late this morning.
Email	Audrey, how about moving the meeting to 8:30? We will have to leave here by 9:35 or so to get a seat at the employee meeting. Kim	Okay, let ' s move Steve ' s Strategy Meeting to 8 : 30a on the 23rd. Please adjust your calendars accordingly. adr Audrey D. Robertson

Table 3: Examples from ECB+ and SC (emails). The same color denotes coreference. Emails are from the same thread.

Improving span representations: Current span representations used in these models lack contextual information. Such contextual information is important for email (e.g. what conversation an email is occurring in). Refining span representations by incorporating whole document contexts or speaker information is an important direction forward.

Word Knowledge & Reasoning: Lastly, one common challenge in current NLU system is the lack of world knowledge. This also holds true for coreference resolution [17]. General information that most parties in the conversation or audiences would recognize will be missing in the email itself. Thus, it may be the case that, in particular for long form conversations, a background knowledge base is a prerequiste for good performance. Such a knowledge base might not be in the form of a knowledge graph but could be in the form of background documents.

7 CONCLUSION

In this paper, we investigated the performance of the state-of-the-art for cross-document coreference resolution on email. We have shown that email is a challenging domain for existing deep learning models. Based on these results and a qualitative analysis, we have identified six paths forward to improve performance in this context. More broadly, we believe understanding these challenges is a first step forward in helping to improve knowledge capture from long form conversations.

Acknowledgement: Funding for this research comes from the Dutch Research Council (NWO) through grant MVI.19.032.

REFERENCES

[1] Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2020. Streamlining Cross-Document Coreference Resolution: Evaluation and Modeling. abs/2009.11032 (2020).

[2] Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In *LREC 2014*.

[3] Parag Pravin Dakle, Takshak Desai, and Dan Moldovan. 2020. A Study on Entity Resolution for Email Conversations. In *Proceedings of the 12th Language Resources and Evaluation Conference*.

[4] Parag Pravin Dakle and Dan I. Moldovan. 2021. CEREC: A Corpus for Entity Resolution in Email Conversations. (2021). arXiv:2105.10606

[5] André Ferreira Cruz, Gil Rocha, and Henrique Lopes Cardoso. 2020. Coreference Resolution: Toward End-to-End and Cross-Lingual Systems. *Information* (2020).

[6] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019. SpanBERT: Improving Pre-training by Representing and Predicting Spans. abs/1907.10529 (2019).

[7] Mandar Joshi, Omer Levy, Daniel S. Weld, and Luke Zettlemoyer. 2019. BERT for Coreference Resolution: Baselines and Analysis. abs/1908.09091 (2019).

[8] Dan Jurafsky and James H. Martin. 2020. *Speech and language processing: 3rd Edition.* Pearson Prentice Hall.

[9] Bryan Klimt and Yiming Yang. 2004. The Enron Corpus: A New Dataset for Email Classification Research. In *ECML*, Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi (Eds.).

[10] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end Neural Coreference Resolution. abs/1707.07045 (2017).

[11] Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-Order Coreference Resolution with Coarse-to-Fine Inference. In *NAACL 2018*.

[12] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. abs/1907.11692 (2019).

[13] James Mayfield, David Alexander, Bonnie J Dorr, et al. 2009. Cross-Document Coreference Resolution: A Key Technology for Learning by Reading.. In *AAAI Spring Symposium: Learning by Reading and Learning to Read.*

[14] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *ACL 2019*.

[15] Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. *NeurIPS 2016* (2016).

[16] Marco Rospocher, Marieke van Erp, Piek Vossen, Antske Fokkens, Itziar Aldabe, German Rigau, Aitor Soroa, Thomas Ploeger, and Tessel Bogaard. 2016. Building event-centric knowledge graphs from news. *Journal of Web Semantics* (2016).

[17] Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2020. Anaphora and coreference resolution: A review. *Information Fusion* (2020).

[18] Nan Wang, Yan Song, and Fei Xia. 2020. Studying Challenges in Medical Conversation with Structured Annotation. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations.*

[19] Liyan Xu and Jinho D. Choi. 2020. Revealing the Myth of Higher-Order Inference in Coreference Resolution. In *Proceedings of the 2020 EMNLP*. Association for Computational Linguistics, Online, 8527–8533. https://doi.org/10.18653/v1/2020.emnlp-main.686

[20] Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *LREC 2017* (2017).