Assigning people to tasks identified in email: The EPA dataset for addressee tagging for detected task intent

Revanth Rameshkumar, Peter Bailey, Abhishek Jha, Chris Quirk
Microsoft, USA
{reramesh, pbailey, abjha, chrisq}@microsoft.com

Abstract

We describe the Enron People Assignment (EPA) dataset, in which tasks that are described in emails are associated with the person(s) responsible for carrying out these tasks. We identify tasks and the responsible people in the Enron email dataset. We define evaluation methods for this challenge and report scores for our model and naïve baselines. The resulting model enables a user experience operating within a commercial email service: given a person and a task, it determines if the person should be notified of the task.

1 Introduction

The initial motivation for our dataset[1] is to enable development of a commercial email service that helps individuals track tasks that are assigned or that they have agreed to perform. To that end, tasks are identified automatically from email text; when such an email is sent or received by an individual, they can be notified or reminded of any resulting tasks should they be responsible for carrying them out. Thus, for the commercial email service there are two parts to the problem: (a) detecting tasks from text, and (b) associating them to specific individuals given a list of affiliated people. The Sender, To and Cc list from the email provide a mostly comprehensive list of individuals. The latter step of selecting responsible individuals is an example of addressee tagging, also referred to as addressee recognition (Traum, 2003).

[1] http://aka.ms/epadataset

> *"Hi Anna, thanks for your work on the sales analysis last week. Can you and Brad complete a draft by Friday please. Thanks, Caira"*

Figure 1. No explicit mention of person.

> *"Hi, thanks for your work on the sales analysis last week. Can you and Brad complete a draft by Friday please. Thanks, Caira*
> *------ Original Message ------*
> *Sent by: Anna Anna@address.com*
> *…*
> *Hi Caira, How are things?*
> *…"*

Figure 2. No explicit mention of person.

For example, in Figure 1, the task is to complete a draft report, and the people responsible are Anna and Brad. However, addressee tagging is needed to match the "you" in the second sentence to "Anna".

As this example demonstrates, often more than one person is responsible for carrying out a task; the task notification must be provided to each responsible party who recieves the email, though not to others who receive the email but are not responsible for the task.

There may be no explicit mention of the person responsible, as in Figure 2. In this example, the context is found not in the user-typed text, but rather in the email client-generated metadata. To complicate matters, the text may use implicit second party references. Consider: "Please complete a draft by Friday". Here, the imperative "complete" has an implicit "you" that can refer to either singular or plural recipients.

Finally, there exist cases in which a task intent is detected, but no one in the To/Cc list is responsible. For example: "Brad will complete the draft report" has a task intent, but if Brad is not in the list of available individuals, we should not erroneously assign one of the recipients.

Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text, pages 28–32
Brussels, Belgium, Nov 1, 2018.

In the commercial email service, both the task and the people responsible must be identified. While the identification of tasks is an interesting problem, including it as part of the challenge would add an additional source of noise to require both the identification of a task and identification of the responsible person or people. Thus, in this dataset, we simplify the problem by providing the set of tasks already extracted from the emails.

2 Dataset and Challenge Description

The dataset consists of a set of tasks in email and the associated people. Within each email, a task is indicated by a special <mark> tag (one per HIT); a set of email recipients (one or more per email) is also provided. The subset of recipients (possibly empty) who are responsible for the marked task is also provided. Each recipient is identified by their email address. Sometimes recipients are email groups("*some_group"*@enron.com); they are referred to implicitly by the sender of the email.

2.1 Enron – Background Email Corpus

The Enron email dataset (Cohen, 2015) was used as a source of tasks described by users in the context of email. In this corpus, most tasks are associated with professional work activities encountered by information workers, but the challenge is emblematic of more general classes of interaction among different groups of people.

The individual emails were pre-processed to produce a standardized format, including email addresses of senders and recipients, subject line, and the textual body of the email.

2.2 Extracted Tasks

A subset of the emails was identified as having tasks: one or more people were requested by the email sender to carry out a specific task. Each task is represented in the dataset as a single sentence from the body of the email, using a basic sentence separation algorithm. As compared to other email datasets, Enron emails are more difficult to segment into sentences due to email formatting. We did not attempt to manually clean "noise" from the sentence segmentation process. The data reflects practical issues when processing email.

The entire email thread is also included as sometimes the prior messages in the thread are helpful in identifying responsible individuals for the task. The prior thread can also contain valuable metadata such as the sender of the previous email.

Task Sentence	**Action**
Can you please send me the document?	Send *[a document]*
Please handle this for John	Handle *[this]*
Please prepare a draft of the letter.	Prepare *[a draft…]*

Table 1. Example tasks and actions

The definition of a task is, broadly, any user intent that requires some explicit subsequent action by one or more individuals. Examples are shown in Table 1.

2.3 Identifying Responsible People

For each task, a set of candidate people is derived from the Sender, To, and Cc email addresses associated with the email. Then, for each person, the challenge is to decide whether that person is responsible for the task that has been identified from the email thread, given both their name and email address. The task is thus reduced to a series of binary decisions, hence we can evaluate using standard binary classification metrics.

2.4 Label Creation Process

In the annotation application, the entire thread is shown with the detected task highlighted in-line. The only preprocessing done on the raw text was to replace
 HTML tags with newlines and replace tabs with spaces, for better results with our production HTML sentence separation algorithm.

All recipients (with the available name and email address information) are shown beside the email along with the two options of 'sender of email' and 'no-one'. The annotators can choose any combination of the 'sender of email' and the recipients; or can choose 'no-one' only.

The annotators were from a *managed* crowd-worker group; they were able to ask us questions, and we could give them feedback on their performance. They were first asked to read the guidelines and complete a qualification task with a passing score of 100%, after an unlimited number of attempts. Generally, the qualification task aims to disqualify crowd-sourced annotaters that are spamming the task or do not understand the task; because our annotator pool is managed, this qualification task also serves as a training tool.

	Perfect Agreement	α
Avocado	84.51 %	0.7854
Enron	71.07%	0.6123

Table 2. Rate of perfect agreement and reliability.

Once the annotators are ready to start the main task, each HIT (here, a single task in one email) is given to three annotators for three independent annotations. A HIT is considered universally agreed upon if all judges agree on all recipients.

We validated the annotations by manually reviewing samples and gave feedback where we found judgements lacking. We also went through several iterations of the task and guidelines to incorporate new feedback, with the dataset using the final iteration of guidelines. Please refer to supplementary material for annotation instructions and pictures to replicate our annotation process on other email (or other) corpora.

2.5 Dataset Analysis

In this section, we provide qualitative and quantitative analysis of the dataset. We also compare the dataset being released to a similarly created dataset from the Avocado email corpus. We cannot release the Avocado version of the dataset due to licensing restrictions.

We use two different agreement calculations since the task seems to be surprisingly subjective and noisy, even after multiple rounds of annotator training. The first is a perfect agreement metric where we simply calculate the number of universally agreed upon label for each (*email*, *task*, *recipient*) tuple (from a total of 15,649 tuples). We found there is a 71.07% perfect agreement rate on the recipient level in the dataset.

To understand rates of inter-annotator agreement better, we calculated Krippendorff's alpha (α), a general and robust reliability measure (Krippendorff, 2004). Our α value is 0.6123. When interpreting magnitude of agreement measures, Krippendorf suggests $\alpha \geq 0.800$, and the threshold for tentative conclusions at 0.667. However, he goes on to say that there is no "magic number" other than perhaps a perfect consensus, and the appropriate α must be determined through experimentation and empirical evidence. In this regard we are still determining an acceptable α for the production scenario. Theoretically the best α would be 1.0, but as we have seen, if we take only data with perfect consensus, we lose up to 28.93% of the collected data (many of which still have a majority consensus). In Table 2, we compare these results on Enron with corresponding annotations over the Avocado dataset (Oard, 2015).

# of unique emails	5998
# of tasks	6300
# of unique recipients in dataset	3460
# of emails with multiple recipients	2923

Table 3. Dataset statistics.

Distribution of # of recipients assigned to task (label = 1)	mean = 1.06 variance = 1.36 range = [0,6]
Distribution of # of recipients not assigned to task (label = 0)	mean = 1.50 variance = 1.60 range = [1,6]

Table 4. Recipient distribution statistics.

As we can see, the agreement and reliability of the Avocado set is substantially higher. When asking the managed annotators if they felt there was any difference between the two sets, and by looking at the data ourselves, the biggest differences seem to be:

1. Avocado has cleaner formatting.
2. Avocado formatting is more consistent, and the annotators find it easier to parse.
3. Avocado sentence separation is cleaner; due to simpler formatting and line breaks.

Future work on the Enron dataset may include additional pre-processing to place less burden on the annotators.

In addition to the agreement and reliability metrics, we have calculated several other statistics in the universally agreed annotated Enron data (Table 3). Similarly, for *email*+*task* combinations with multiple recipients, we report basic statistics on distribution of recipients in Table 4.

3 Evaluation for Task

In the production scenario, performance is measured by the precision and recall of task assignment to a recipient on the recipient list. The other two metrics we looked at were precision and recall of single recipient vs multi recipient emails, and the distribution of precision and recall for each email. The calculation of the precision and recall is based on the simple binary label assigned to the (*email*, *task*, *recipient*) tuple.

4 Model and Performance

4.1 Baselines

The baselines are detailed in Table 5. We consider the naïve baselines of assuming every person is responsible for the task in the single recipient and multi-recipient case. We also have the baseline of assigning a person to the task with the mean probability of a recipient being responsible for the task. This probability is lower than 1.0 in the single recipient case because sometimes 'NOBODY' is responsible for the task.

Finally, we provide the baseline of a model trained on the Avocado dataset (our first dataset, and the model in production) on the Enron dataset. This is an interesting result: an Avocado trained model and evaluated on an Avocado blind set yielded a P/R of 0.9/0.9. It also performs reasonably well on a donated sample of real user emails; yet performs relatively poorly on the new Enron dataset. All baselines are calculated on (*email*, *task*, *recipient*) tuples with total consensus.

Single Recipient

Prediction Strategy	P	R	F1
Every recipient	0.6730	1.0000	0.8045
$\bar{x}$ (0.6674)	0.6589	0.6503	0.6545
Avocado model	0.6917	0.8927	0.7795

Multiple Recipients

Prediction Strategy	P	R	F1
Every recipient	0.4435	1.0000	0.6145
$\bar{x}$ (0.4289)	0.4448	0.4280	0.4362
Avocado model	0.6169	0.7021	0.6567

Table 5. Baseline performance for single and multiple recipients.

4.2 Experimental Model

To train the model, we are currently using the data collected from the Avocado training set. We have trained on 3872 emails, and we took the majority consensus HIT (otherwise random). In the future we can try to incorporate annotator reliability metrics (Rehbein, 2017) to allow us to filter for more data.

The model is trained using logistic regression with a set of handcrafted features (described in supplementary material). The best feature contributions come from token replacement and encoding out-of-sentence token information.

5 Future work

We plan to improve pre-processing, in hopes of raising the inter-annotator agreement on the Enron set to at least match the Avocado set. Also, the deictic nature of this task can be extended from addressee assignment to time and location assignment. A temporal expression and location tagger can be used to build the set of assignable entities to the extracted task, and we could contrast the application of explicit linguistic features or implicitly learned features developed from addressee assignment to the task of assigning time and location.

6 Related Work

Many addressee detection methods have been developed in dialogue-based domains, such as in the use of the AMI meeting corpus for addressee detection (Akker and Traum, 2009). The corpus is a set of 14 meetings in which utterances were captured, and "important" utterances were labeled with the addressee. The addressee is labeled as whole group or one of four individuals. The manual annotation effort in that effort also seems to exhibit an α value below 0.8.

Purver (2006) used the ICSI and ISL meeting corpora to label task owner via utterance classification, in addition to other utterance labels. As we have, Purver et al. noticed that labels for owner and other task properties might be derived from context nearby the utterance containing the actual task. Though they report a kappa score of 0.77, they also note that their model performed worst on owner classification.

Kalia et al. attempted to detect of commitments from a subset of the Enron dataset (Kalia, 2013). This subset came from exchanges involving a specific user, and the focus was on the task extraction. There was no specific effort to label task owner. The inter-annotator metric was a kappa score of 0.83 (combined with their chat dataset). We speculate that using a more specific task format and not using context led to increased agreement.

7 Conclusion

We introduce the Enron People Assignment dataset, containing addressee assignment annotations, for 15,649 (*email*, *task*, *recipient*) tuples, for the noisy task of assigning the proper recipient to an extracted task. We analyzed the

dataset, calculated reliability and agreement metrics, and provided baselines for people assignment task. Our experiments show that annotation and model performance vary significantly across datasets, and that there is substantial room for improvement when modeling people assignment in the email domain alone. Our broader goal in releasing this task and dataset is to motivate researchers to develop new methods to process and model noisy, yet rich, text.

Acknowledgments

We wish to thank the Substrate Query and Intelligence team in Microsoft AI&R, whose great work allowed us to embark on this project. We are also grateful to the reviewers for their time.

References

Rieks op den Akker and David Traum. (2009). A comparison of addressee detection methods for multiparty conversations. In *Workshop on the Semantics and Pragmatics of Dialogue*.

William W. Cohen. (2015). Enron Email Dataset. URL: `www.cs.cmu.edu/~./enron/`.

Douglas Oard, et al. (2015) Avocado Research Email Collection LDC2015T03. DVD. Philadelphia: Linguistic Data Consortium.

Antoine, Jean-Yves, Jeanne Villaneau, and Anaïs Lefeuvre. (2014, April). Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. In EACL 2014 (pp. 10-p).

Ning Gao, Gregory Sell, Douglas W. Oard, and Mark Dredze. (2017, December). Leveraging side information for speaker identification with the Enron conversational telephone speech collection. In *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE* (pp. 577-583). IEEE

Anup Kalia, Hamid Reza Motahari Nezhad, Claudio Bartolini, and Munindar Sing (2013). Identifying business tasks and commitments from email and chat conversations. *Citeseer, Tech. Rep.*

Klauss Krippendorff. (2004). Content analysis: An introduction to its methodology. Second Edition. Thousand Oaks, CA: Sage.

Andrew F. Hayes and Klaus Krippendorff. (2007). Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, *1*(1), 77-89.

Zhao Meng, Lili Mou, and Zhi Jin. (2017, November). Hierarchical RNN with Static Sentence-Level Attention for Text-Based Speaker Change Detection. In *Proc. 2017 ACM on Conference on Information and Knowledge Management* (pp. 2203-2206). ACM.

Michael Muller, Casey Dugan, Michael Brenndoerfer, Megan Monroe, and Werner Geyer. (2017, February). What Did I Ask You to Do, by When, and for Whom?: Passion and Compassion in Request Management. In *Proc. 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 1009-1023). ACM.

Matthew Purver, Patrick Ehlen, and John Niekrasz. (2006, May). Detecting action items in multi-party meetings: Annotation and initial experiments. In *Proc. Third International Conference on Machine Learning for Multimodal Interaction* (pp. 200–211). Springer-Verlag.

David Traum. (2003, July). Issues in multiparty dialogues. In *Workshop on Agent Communication Languages* (pp. 201-211). Springer, Berlin, Heidelberg.

Rehbein, Ines, and Josef Ruppenhofer. (2017). Detecting annotation noise in automatically labelled data. In *Proc. 55th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (Vol. 1, pp. 1160-1170).

Gupta, Surabhi, John Niekrasz, Matthew Purver, and Daniel Jurafsky. (2007, September). Resolving "you" in multiparty dialog. In *Proc*. 8th SIGdial Workshop on Discourse and Dialogue (pp. 227-30).

Jovanovic, Natasa, Rieks op den Akker, and Anton Nijholt. (2006). A corpus for studying addressing behaviour in multi-party dialogues. Language Resources and Evaluation, 40(1), 5-23.

Webb, Nick, Mark Hepple, and Yorick Wilks. (2005, July). Dialogue act classification based on intra-utterance features. In *Proc. AAAI Workshop on Spoken Language Understanding* (Vol. 4, p. 5).

Yang Liu, Kun Han, Zhao Tan, and Yun Lei. (2017). Using Context Information for Dialog Act Classification in DNN Framework. In *Proc. 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2170-2178).

Jacob Cohen (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological bulletin, 70(4), 213.

Oluwasanmi O. Koyejo, Nagarajan Natarajan, Pradeep K. Ravikumar, and Inderjit S. Dhillon. (2014). Consistent binary classification with generalized performance metrics. In *Proc. Advances in Neural Information Processing Systems* (pp. 2744-2752).