Annotation and Classification of an Email Importance Corpus

Fan Zhang
Computer Science Department
University of Pittsburgh
Pittsburgh, PA 15260
zhangfan@cs.pitt.edu

Kui Xu
Research and Technology Center
Robert Bosch LLC
Palo Alto, CA 94304
Kui.Xu2@us.bosch.com

Abstract

This paper presents an email importance corpus annotated through Amazon Mechanical Turk (AMT). Annotators annotate the email content type and email importance for three levels of hierarchy (senior manager, middle manager and employee). Each email is annotated by 5 turkers. Agreement study shows that the agreed AMT annotations are close to the expert annotations. The annotated dataset demonstrates difference in proportions of content type between different levels. An email importance prediction system is trained on the dataset and identifies the unimportant emails at minimum 0.55 precision with only text-based features.

1 Introduction

It is common that people receive tens or hundreds of emails everyday. Reading and managing all these emails consume significant time and attention. Many efforts have been made to address the email overload problem. There are studies modeling the email importance and the recipients' actions in order to help with the user's interaction with emails (Dabbish and Kraut, 2006; Dabbish et al., 2005). Meanwhile, there are NLP studies on spam message filtering, email intention classification, and priority email selection to reduce the number of emails to read (Schneider, 2003; Cohen et al., 2004; Jeong et al., 2009; Dredze et al., 2009). In our project, we intend to build an email briefing system which extracts and summarizes important email information for the users.

However, we believe there are critical components missing from the current research work. First, to the extent of our knowledge, there is little public email corpus with email importance labeled. Most of the prior works were either based on surveys or private commercial data (Dabbish and Kraut, 2006; Aberdeen et al., 2010). Second, little attention has been paid to study the difference of emails received by people at different levels of hierarchy. Third, most of the prior works chose the user's action to the email (e.g. replies, opens) as the indicator of email importance. However, we argue that the user action does not necessarily indicate the importance of the email. For example, a work-related reminder email can be more important than a regular social greeting email. However, a user is more likely to reply to the later and keep the information of the former in mind. Specifically for the goal of our email briefing system, importance decided upon the user's action is insufficient.

This paper proposes to annotate email importance on the Enron email corpus (Klimt and Yang, 2004). Emails are grouped according to the recipient's levels of hierarchy. The importance of an email is annotated not only according to the user's action but also according to the importance of the information contained in the email. The content type of the emails are also annotated for the email importance study. Section 3 describe the annotation and analysis of the dataset. Section 4 describes our email importance prediction system trained on the annotated corpus.

2 Related work

The most relevant work is the email corpus annotated by Dredze et al. (Dredze et al., 2008a; Dredze et al., 2008b). 2391 emails from inboxes of 4 volunteers were included. Each volunteer manually annotated whether their own emails need to be replied or not. The annotations are reliable as they come from the emails' owners. However, it lacks diversity in the user distribution with only 4 volunteers. Also, whether an email gets response or not does not always indicate its importance. While commercial products such as Gmail Priority Inbox (Aberdeen et al., 2010) has a better cover-

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 651–656,
Beijing, China, July 26-31, 2015.

age of users and decides the importance of emails upon more factors[1], it is unlikely to have their data accessible to public due to user privacy concerns.

The Enron corpus is a public email corpus widely researched (Klimt and Yang, 2004). Lampert et al. (2010) annotated whether an email contains action request or not based on the agreed annotations of three annotators. We followed similar ideas and labeled the email importance and content type with the agreed Amazon Mechanical Turk annotations. Emails are selected from Enron employees at different levels of hierarchy and their importance are labeled according to the importance of their content. While our corpus can be less reliable without the annotations from the emails' real recipients, it is more diverse and has better descriptions of email importance.

3 Data annotation

3.1 Annotation scheme

Annotators are required to select the importance of the email from three levels: *Not important*, *Normal* and *Important*. *Not important* emails contain little useful information and require no action from the recipient. It can be junk emails missed by the spam filter or social greeting emails that do not require response from the recipient. *Important* emails either contain very important information to the recipient or contain urgent issues that require immediate action (e.g. change of meeting time/place). *Normal* emails contain less important information or contain less urgent issues than *Important* emails. For example, emails discussing about plans of social events after work would typically be categorized as *Normal*.

We also annotate the email content type as it reveals the semantic information contained in the emails. There are a variety of email content type definitions (Jabbari et al., 2006; Goldstein et al., 2006; Dabbish et al., 2005). We choose Dabbish et al.'s definition for our work. Eight categories are included: *Action Request*, *Info Request*, *Info Attachment*, *Status Update*, *Scheduling*, *Reminder*, *Social*, and *Other*. While an email can contain more than one type of content, annotators are required to select one primary type.

[1]Including user actions and action time, the user actions not only include the *Reply* action but also includes actions such as *opens*, *manual corrections*, etc.

3.2 Annotation with AMT

Amazon Mechanical Turk is widely used in data annotation (Lawson et al., 2010; Marge et al., 2010). It is typically reliable for simple tasks. Observing the fact that it takes little time for a user to decide an email's importance, we choose AMT to do the annotations and manage to reduce the annotation noise through redundant annotation.

Creamer et al. categorized the employees of the Enron dataset to 4 groups: senior managers, middle managers, traders and employees[2] (Creamer et al., 2009). We hypothesized that the types of emails received by different groups were different and annotated different groups separately. Based on Creamer et al's work, we identified 23 senior managers with a total of 21728 emails, 20 middle managers with 13779 emails and 17 regular employees with 12137 emails. The trader group was not annotated as it was more specific to Enron. For each group, one batch of 750 assignments (email) was released. The emails were randomly selected from all the group members' received emails (to or cc'ed). Turkers were presented with all details available in the Enron dataset, including subject, sender, recipients, cclist, date and the content (with history of forwards and replies). Turkers were required to make their choices as they were in the position.[3] Each assignment was annotated by 5 turkers at the rate of $0.06 per Turker assignment. The email type and the email importance are decided according to the majority votes. If an email has 3 agreed votes or higher, we call this email **agreed**. Table 1 demonstrates the average time per assignment (Time), the effectively hourly rate (Ehr), the number of emails with message type agreed (#TypeAgreed), importance agreed (#ImpoAgreed) and both agreed (#AllAgreed). We find that #AllAgreed is close to #TypeAgreed, which indicates a major overlap between the agreed type annotation and the agreed importance annotation.

3.3 Data discussion

In this paper we focus on the *AllAgreed* emails to mitigate the effects of annotation noise. Table 2 demonstrates the contingency table of the corpus.

[2]Senior managers include CEO, presidents, vice presidents, chief risk officer, chief operating officer and managing directors. The other employees at management level are categorized to middle managers

[3]E.g. instruction of the senior manager batch: Imagine you were the CEO/president/vice president/managing director of the company, categorize the emails into the three categories [Not Important], [Normal], [Important].

Level	Time (s)	Ehr ($)	#All	#TypeAgreed	#ImpoAgreed	#AllAgreed
Senior (23)	40	5.400	750	589	656	574
Middle (20)	33	6.545	750	556	622	550
Employee (17)	31	6.968	750	593	643	586

Table 1: AMT annotation results, notice that #AllAgreed is close to #TypeAgreed

	Act.Req	Info.Req	Info	Status	Schedule	Reminder	Social	Other	All
Senior	60	49	255	**57**	43	4	**68**	38	574
Not	0	0	0	0	0	0	33	30	63
Normal	38	37	231	51	37	4	35	8	441
Important	22	12	24	6	6	0	0	0	70
Middle	82	53	261	22	49	0	37	46	550
Not	0	0	1	0	0	0	10	32	43
Normal	64	47	247	22	49	0	27	14	470
Important	18	6	13	0	0	0	0	0	37
Employee	61	65	326	22	29	1	52	30	586
Not	0	0	1	0	0	0	8	26	35
Normal	43	62	315	22	27	1	44	4	518
Important	18	3	10	0	2	0	0	0	33

Table 2: Contingency table of content type and importance of *AllAgreed* emails; bold indicates the proportions of this category is significantly different between groups ($p<0.05$)

A potential issue of the corpus is that the importance of the email is not decided by the real email recipient. To address this concern, we compared the *AllAgreed* results with the annotations from an expert annotator. 50 emails were randomly selected from *AllAgreed* emails for each level. The annotator was required to check the background of each recipient (e.g. the recipient's position in the company at the time, his/her department information and the projects he/she was involved in if these information were available online) and judge the relationship between the email's contacts before annotation (e.g. if the contact is a family member or a close friend of the recipient). Agreement study shows a Kappa score of 0.7970 for the senior manager level, 0.6420 for the middle manager level and 0.7845 for the employee level. It demonstrates that the agreed Turker annotations are as reliable as well-prepared expert annotations.

We first tested whether the content type proportions were significantly different between different levels of hierarchy. Recipients with more than 20 emails sampled were selected. A vector of content type proportions was built for each recipient on his/her sampled emails. Then we applied multivariate analysis of variance (MANOVA) to test the difference in the means of the vectors between levels[4]. We found that there were significant differences in proportions of status update ($p=0.042$) and social emails ($p=0.035$). This agrees with the impression that the senior managers spend more time on project management and social relationship development. Following the same approach, we tested whether there were significant differences in importance proportions between levels. However, no significant difference was found while we can observe a higher portion of *Important* emails in the *Senior* group in Table 2. In the next section, we further investigate the relationship between content type and message importance using the content type as a baseline feature in email importance prediction.

4 Email importance prediction

In this section we present a preliminary study of automatic email importance prediction. Two baselines are compared, including a *Majority* baseline where the most frequent class is chosen and a *Type* baseline where the only feature used for classification is the email content type.

[4]We cannot use Chi-square to test the difference between groups directly on Table 2 as the emails sampled do not satisfy the independence consumption if they come from the same recipient

Features	Acc	Kappa	P(U)	R(I)
Sr. Mgrs				
Majority	76.83	0	0	0
Type	68.78	37.93	58.76	44.81
Text	76.34	26.96	71.83∗	14.67†
Text+Type	78.43	33.80	**75.99**∗	12.13†
Mgrs				
Majority	85.45	0	0	0
Type	69.81	32.75	50.47	49.80
Text	87.09	26.64	54.67	4.17†
Text+Type	88.55	36.42	63.80∗	7.59†
Emp				
Majority	88.39	0	0	0
Type	80.34	38.63	40.21	45.12
Text	88.83	30.98	63.83∗	1.67†
Text+Type	89.16	36.71	72.50∗	1.67†

Table 3: Results of Experiment 1; ∗ indicates significantly better than the Type baseline; † indicates significantly worse than the Type baseline; bold indicates better than all other methods. With only text-based features, the system achieves at least 54.67 precision in identifying unimportant emails.

Groups	Acc	Kappa	P(U)	R(I)
Sr. Mgrs	77.70	19.24	65.22	10.00
Mgrs	83.27	30.03	61.90	2.70
Emp	83.10	33.89	46.94	33.33

Table 4: Cross-group results of Experiment 2

4.1 Feature extraction

While prior works have pointed out that the social features such as contacting frequency are related to the user's action on emails (Lampert et al., 2010; Dredze et al., 2008a), in this paper we only focus on features that can be extracted from text.

N-gram features Binary unigram features are extracted from the email subject and the email content separately. Stop words are not filtered as they might also hint the email importance.

Part-of-speech tags According to our observation, the work-related emails have more content words than greeting emails. Thus, POS tag features are extracted from the email content, including the total numbers of POS tags in the text and the average numbers of tags in each sentence. [5]

Length features We observe that work-related emails tend to be more succinct than unimportant emails such as advertisements. Thus, length features are extracted including the length of the email subject and email content, and the average length of sentences in the email content.

Content features Inspired by prior works (Lampert et al., 2010; Dredze et al., 2008a), features that provide hints of the email content are extracted, including the number of question marks, date information and capitalized words, etc.

4.2 Experiments

We treat our task as a multi-class classification problem. We test classifications within-level and cross-level with only text-based features.

Experiment 1 Each level is tested with 10-fold cross-validation. SVM of the Weka toolkit (Hall et al., 2009) is chosen as the classifier. To address the data imbalance problem, the minority classes of the training data are oversampled with the Weka SMOTE package (Chawla et al., 2002). The parameters of SMOTE are decided according to the class distribution of the training data.

Experiment 2 The classifiers are trained on two levels and tested on the other level. Again, SVM is chosen as the model and SMOTE is used to oversample the training data.

4.3 Evaluation

Kappa[6] and *accuracy* are chosen to evaluate the overall performance in prediction. For our email briefing task specifically, *precision* in unimportant email prediction P(U) (avoid the false recognition of unimportant emails) and *recall* in important email prediction R(I) (cover as many important emails as possible) are evaluated. Paired t-tests are utilized to compare whether there are significant differences in performance ($p < 0.05$).

As demonstrated in Table 3, the text-based features are useful for the prediction of unimportant email classification but not as useful for the recognition of important emails. It also shows that the content type is an important indicator of the email's importance. While the content type is not always accessible in real life settings, the results demonstrate the necessity of extracting semantic information for email importance prediction. In Table 4, precision of unimportant email prediction

[5]The Part-of-speech (POS) tags are tagged with the Stanford CoreNLP toolkit (Manning et al., 2014; Toutanova et al., 2003), containing 36 POS tags as defined in the Penn Treebank annotation.

[6]The agreement between the system and the majority labels from the Mechanical Turk

is higher on the manager levels but lower on the employee level. This indicates a potential difference of email features between the manager levels and the employee level.

5 Conclusion and future work

In this paper we present an email importance corpus collected through AMT. The dataset focuses on the importance of the information contained in the email instead of the email recipient's action. The content type of the email is also annotated and we find differences in content type proportions between different levels of hierarchy. Experiments demonstrate that the content type is an important indicator of email importance. The system based on only text-based features identifies unimportant emails at minimum 0.5467 precision.

Agreement study shows that the agreed Turker annotations are as good as annotations of well-prepared expert annotators. We plan to increase the size of our dataset through AMT. We expect the dataset to be helpful for studies on email overload problems. Meanwhile, we are aware that the current corpus lacks social and personal information. We believe features regarding such information (e.g. the recipient's email history with the contact, the recipient's personal preference in categorizing emails, etc.) should also be incorporated for importance prediction.

Acknowledgments

We would like to thank Zhe Feng, Doo Soon Kim, Lin Zhao, Sai Sudharsanan and other employees of the Bosch RTC 1.2 group for their helpful feedback, Prof. Diane Litman for her instructions to the first author and all the anonymous reviewers for their suggestions.

This research is supported by the Research and Technology Center of Robert Bosch LLC.

References

Douglas Aberdeen, Ondrej Pacovsky, and Andrew Slater. 2010. The learning behind gmail priority inbox. In *LCCC: NIPS 2010 Workshop on Learning on Cores, Clusters and Clouds*.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16(1):321–357.

William W. Cohen, Vitor R. Carvalho, and Tom M. Mitchell. 2004. Learning to classify email into "speech acts". In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 309–316, Barcelona, Spain, July. Association for Computational Linguistics.

Germán Creamer, Ryan Rowe, Shlomo Hershkop, and Salvatore J Stolfo. 2009. Segmentation and automated social hierarchy detection through email network analysis. In *Advances in Web Mining and Web Usage Analysis*, pages 40–58. Springer.

Laura A Dabbish and Robert E Kraut. 2006. Email overload at work: an analysis of factors associated with email strain. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 431–440. ACM.

Laura A Dabbish, Robert E Kraut, Susan Fussell, and Sara Kiesler. 2005. Understanding email use: predicting action on a message. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 691–700. ACM.

Mark Dredze, Tova Brooks, Josh Carroll, Joshua Magarick, John Blitzer, and Fernando Pereira. 2008a. Intelligent email: reply and attachment prediction. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 321–324. ACM.

Mark Dredze, Hanna M Wallach, Danny Puller, Tova Brooks, Josh Carroll, Joshua Magarick, John Blitzer, Fernando Pereira, et al. 2008b. Intelligent email: Aiding users with ai. In *AAAI*, pages 1524–1527.

Mark Dredze, Bill N Schilit, and Peter Norvig. 2009. Suggesting email view filters for triage and search. In *IJCAI*, pages 1414–1419.

Jade Goldstein, Andres Kwasinksi, Paul Kingsbury, Roberta Evans Sabin, and Albert McDowell. 2006. Annotating subsets of the enron email corpus. In *CEAS*. Citeseer.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Sanaz Jabbari, Ben Allison, David Guthrie, and Louise Guthrie. 2006. Towards the orwellian nightmare: separation of business and personal emails. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 407–411. Association for Computational Linguistics.

Minwoo Jeong, Chin-Yew Lin, and Gary Geunbae Lee. 2009. Semi-supervised speech act recognition in emails and forums. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1250–1259. Association for Computational Linguistics.

Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *Machine learning: ECML 2004*, pages 217–226. Springer.

Andrew Lampert, Robert Dale, and Cecile Paris. 2010. Detecting emails containing requests for action. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 984–992. Association for Computational Linguistics.

Nolan Lawson, Kevin Eustice, Mike Perkowitz, and Meliha Yetisgen-Yildiz. 2010. Annotating large email datasets for named entity recognition with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 71–79. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Matthew Marge, Satanjeev Banerjee, and Alexander I Rudnicky. 2010. Using the amazon mechanical turk to transcribe and annotate meeting speech for extractive summarization. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 99–107. Association for Computational Linguistics.

Karl-Michael Schneider. 2003. A comparison of event models for naive bayes anti-spam e-mail filtering. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 307–314. Association for Computational Linguistics.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.