Information Sciences 358–359 (2016) 1–17

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

# Assessing e-mail intent and tasks in e-mail messages

M. Sappelli[a,b,*], G. Pasi[c,*], S. Verberne[b], M. de Boer[a,b], W. Kraaij[a,b]

[a] TNO, Anna van Buerenplein 1, Den Haag, The Netherlands
[b] Radboud University, Toernooiveld 300, Nijmegen, The Netherlands
[c] Universit di Milano Bicocca, Italy

ARTICLE INFO

*Article history:*
Received 28 July 2015
Revised 4 February 2016
Accepted 4 March 2016
Available online 12 March 2016

*Keywords:*
E-mail intent
E-mail annotation scheme
Task-based e-mail categorization
Human annotation

ABSTRACT

In this paper, we analyze corporate e-mail messages as a medium to convey work tasks. Research indicates that categorization of e-mail could alleviate the common problem of information overload. Although e-mail clients provide possibilities of e-mail categorization, not many users spend effort on proper e-mail management. Since e-mail clients are often used for task management, we argue that intent- and task-based categorizations might be what is missing from current systems.

We propose a taxonomy of tasks that are expressed through e-mail messages. With this taxonomy, we manually annotated two e-mail datasets (Enron and Avocado), and evaluated the validity of the dimensions in the taxonomy. Furthermore, we investigated the potential for automatic e-mail classification in a machine learning experiment.

We found that approximately half of the corporate e-mail messages contain at least one task, mostly informational or procedural in nature. We show that automatic detection of the number of tasks in an e-mail message is possible with 71% accuracy. One important finding is that it is possible to use the e-mails from one company to train a classifier to classify e-mails from another company. Detecting how many tasks a message contains, whether a reply is expected, or what the spatial and time sensitivity of such a task is, can help in providing a more detailed priority estimation of the message for the recipient. Such a priority-based categorization can support knowledge workers in their battle against e-mail overload.

## 1. Introduction

In the project SWELL[1] we aim to develop ICT applications that minimize the risk of burn-out and improve the well-being of employees. A large source of stress at work originates from information overload, and more specifically e-mail overload [2,13]. Whittaker and Sidner [34] believe that this is caused by the misuse of the original purpose of the e-mail system. The authors state that although e-mail was originally developed for the purpose of asynchronous communication, it is currently being used for task management, scheduling and personal archiving as well. This causes cluttered in-boxes and information getting lost in archives.

[*] Corresponding author at: TNO, Anna van Buerenplein 1, Den Haag, The Netherlands.
*E-mail addresses:* maya.sappelli@tno.nl (M. Sappelli), pasi@disco.unimib.it (G. Pasi), s.verberne@cs.ru.nl (S. Verberne), maaike.deboer@tno.nl (M. de Boer), kraaijw@acm.org (W. Kraaij).

[1] http://www.swell-project.net

http://dx.doi.org/10.1016/j.ins.2016.03.002
0020-0255/© 2016 Elsevier Inc. All rights reserved.

Attempts to improve the organization of inboxes include the automatic detection of spam [28], message categorization [3,8,15,22,27,30] and priority estimation [1,11,29]. Complete agents exist that help the user file messages into folders [31]. However, not many of these automated techniques are adopted in current systems and many users do not even use category folders at all [15,22]. The most likely actions users make are splitting personal and work-related e-mail by using separate mailboxes [7] and cleaning e-mails at the end of the day [19]. In general, not many users spend effort on general e-mail management (deleting, moving, flagging) [17]. Nevertheless, research indicates that proper categorizations could alleviate the problem of feeling overloaded [2,4,34]. The fact that categorizations are not used suggests that there may not be a full understanding of what type of categorization is needed to properly support users in the way they use e-mail.

Since e-mail clients are often used for task management [7,34], we believe that intent- and task-based categorizations might be what is missing from current systems. This paper studies the realization of tasks in e-mail messages to better understand what the intent is behind an e-mail. In order to do so we annotate e-mail messages with both their e-mail intent and task intent. By e-mail intent we mean the intent of the sender; why did a person send the message. In that case, the intent refers to a message as a whole. Then, within a message the sender has (either implicitly or explicitly) possibly specified one or more tasks to be undertaken by the receiver. This latter aspect is referred to as the task(s) in the message. In this paper we investigate both the intent of the message and the tasks that are conveyed in the message. Additionally, we investigate how an e-mail conversation between two individuals evolves over time.

This paper makes four contributions compared to the previous literature: First, a taxonomy of tasks that can be found in e-mail messages is proposed. Second, a manual annotation of two datasets is provided; these annotations will be available to the research community to provide new opportunities for the development of (automated) e-mail support systems. Third, we present an initial analysis of how senders convey tasks in e-mail messages. And finally we present some initial results on automated classification for one of the dimensions. We address the following research questions:

1. To what extent do corporate e-mail messages contain tasks?
2. What are the characteristics of tasks in e-mail messages?
3. How do work-related e-mail conversations evolve?
4. Can we determine the number of tasks in a message automatically?

We start this paper with an overview of literature on the analysis of e-mail message content. Then we present the results of a pilot study where we developed our e-mail classification scheme in Section 3. In this study we determine which dimensions of content analysis are reliable for annotation. Additionally, we assess the validity of using annotations by independent assessors. In Section 4, we present the results of a larger-scale annotation study, where we annotate messages from the Enron and Avocado datasets. These datasets originate from a company setting and are likely to be representative of how tasks are conveyed in a work environment. In Section 5 we demonstrate the possibility to use the collected data for automated classification.

## 2. Background literature

In this section we shortly report the literature related to the analysis of e-mail message content. A limitation of the research that addresses the analysis of e-mail messages is the limited availability of public datasets of e-mail messages. The most used publicly available dataset of e-mail messages is the Enron dataset. This is a set of messages that was made public during a legal investigation of the Enron company [20]. It contains over 200,000 messages. Many researchers, however, make use of their own privately collected sets of e-mail messages [1,10,11,21].

Some research into the content of e-mail messages has been directed at the timing of communication. In an interview study, Tyler and Tang [32] investigate the concept of the *responsiveness image* of a person in order to understand what information is conveyed by the timing of email responses. They distinguish *response expectation* (the implicit time the sender gives to the recipient to respond) from *breakdown perception* (the initiation of a follow-up action that occurs when the response expectation time has ended).

This responsiveness image could be seen as a request for attention. Hanrahan et al. [17] analyze responsiveness in a 2-week study by logging user interactions with e-mail and compared these interactions to diary entries of the participants. The authors propose that e-mails can be categorized into 4 groups of requests for attention: ignore, accountable non-answer (engage with message but do not reply), postponed reply and immediate reply. This categorization provides insight in both the timing as well as the type of response that is expected.

Kooti et al. [21] add that the request of attention is not solely based on the contents of a message. They note that there is an effect of load on the replying behavior of people. As users receive more e-mail messages in a day, they will reply to a smaller fraction of messages.

Besides the research centered on replying behavior, another line of research that addresses the issue of identifying e-mail intent is finalized at an analysis of the content of messages. Gains [12] focuses on the language that is used in messages, in particular on aspects related to the pattern and style of a text. He has analyzed messages in a commercial and in an academic setting and found that commercial e-mail messages tend to follow standard written business English, while messages in an academic setting follow a more pseudo-conversational pattern where for example the salutation is absent.

To analyze the message style, Gains [12] uses a classification scheme from business communication described by Ghadessy and Webster [14]. The authors state that there are roughly three types of business communication: informative (give information), requestive (request information) and directive (give instructions). Furthermore Ghadessy and Webster [14] distinguish an initiate and a respond category. These categorizations seem intuitive descriptors related to e-mail intent.

In addition to the business communication categorization, Peterson et al. [26] assessed the formality of e-mail messages in the Enron corpus. They annotated 400 messages on a 4-point scale (very formal, somewhat formal, somewhat informal and very informal). Factors that influenced the formality of the messages were the amount of contact between sender and recipient, whether it was personal or business, the rank difference between sender and recipient, and whether the message contained a request.

In terms of the tasks in e-mail, some previous research adopts categories from speech act theory. Cohen et al. [9] propose to categorize e-mails according to the intent of the sender. They propose to use categories of intent based on speech act. The categories are *meeting*, *deliver*, *commit*, *request*, *amend* and *propose*. They later refine this categorization [6], which is explained in more detail in Section 3.1

In addition, Lampert et al. [23] have conducted several e-mail labeling experiments on Enron data to evaluate reliability of task-based intent assessments. They focus on the speech acts of request and commit. They found that the assessments were more reliable on the message level compared to the sentence level. This suggests that messages should be evaluated as a whole.

Kalia et al. [18] not only describe the identification of tasks based on speech act theory, but also the tracking of tasks. They distinguish the following phases: the creation of a commitment, the discharge of a commitment, the delegation of the commitment and the cancellation of the commitment. Their algorithms require detailed NLP analysis of the message to determine what the subject, object and action is in their tasks. This is necessary to determine whether a task is delegated to another person. Their algorithms were evaluated on a selection of 4161 sentences from the Enron corpus.

There are several disadvantages in the methods described in this section. First of all, many of the annotated data sets that were created with the proposed annotation schemes are not publicly available. This limits the possibility to use the annotations for other purposes, or to compare the annotations between different datasets. Moreover, some of the annotation schemes are focused on only a very targeted aspect of a message, such as the formality of the language that is used. This limits the possibility of using the annotation scheme for another purpose such as, in our case, identifying tasks and thereby supporting knowledge workers. And finally, not all annotation schemes have a clear relation to task and intent. We believe that understanding the intent of email messages and the tasks contained in them is the best approach to support knowledge workers.

In our work we focus on both the intent-based and the task-based categorization of e-mail messages on multiple dimensions. In the next section we describe the dimensions that we take into consideration and assess the reliability and validity of those dimensions in a pilot annotation study with private e-mails.

## 3. Reliability and validity of e-mail annotations

From the background literature we identified several dimensions of email intent categorization, such as response expectation (also referred to as reply expectation), speech act and formality. We start with a pilot study in which we have determined the dimensions that are relevant for the assessment of *e-mail intent* and the understanding of *task conveyance* in e-mail. With *task conveyance* we mean the communication of a task to the recipient of a message. The goal is to create a reliable and valid taxonomy for a task-based classification of e-mail messages. In Section 4 this taxonomy is use to annotate two datasets and in Section 5 we describe the automated classification of messages on one of the dimensions from the taxonomy.

In order to determine which dimensions are relevant for task-based classification of e-mail, we assess the reliability and the validity of candidate annotation dimensions. A reliable dimension is a dimension on which two or more annotators typically agree in their annotations (inter-rater reliability). A valid dimension is a dimension where independent assessors typically agree with the ground truth annotation [33]. The sender of the message determines the ground truth annotation, as his intent is the one we try to assess. The reliability and the validity of the dimensions respectively support the selection of which dimensions we should annotate in our main study, and whether independent assessors are capable of assessing the sender's intent. The reason why we assess this in a pilot study where the messages are not publicly available, is that we need to involve the original senders of the messages to assess the validity of the proposed annotation scheme. We do not have this possibility for the datasets that we use in the main study. This is a limitation, since the pilot study only includes a limited number of e-mail messages because of the labor-intensive nature of the annotation work.

### *3.1. E-mail intent and task classification scheme*

In the selection of the dimensions for our e-mail intent and task classification scheme we have focused on those dimensions that are related to the content of the message, and more specifically the tasks that are conveyed through sending the message. On the one hand these dimensions are related to the message as a whole; what was the intent of the sender, what implicit reason was there for sending the message etc. On the other hand the dimensions are related to explicit tasks that are mentioned in the message; what is the recipient supposed to do after reading the message, how many tasks are mentioned, what is their spatial and time sensitivity and what kind of task is it.

**Table 1**
Dimensions related to the intent of e-mail messages; what was the motivation of the sender to send the message.

| Dimension | Description |
|---|---|
| E-mail acts [6] | What are the two main e-mail acts in the message? This dimension has categorical values, consisting of:<br>*Request*: A request asks (or orders) the recipient to perform some activity. A question is also considered a request (for delivery of information)<br>*Propose*: A propose message proposes a joint activity, i.e., asks the recipient to perform some activity and commits the sender as well, provided the recipient agrees to the request. A typical example is an email suggesting a joint meeting<br>*Commit*: A commit message commits the sender to some future course of action, or confirms the sender's intent to comply with some previously described course of action<br>*Deliver*: A deliver message delivers something, e.g., some information, a PowerPoint presentation, the URL of a website, the answer to a question, a message sent “FYI”, or an opinion<br>*Amend*: An amend message amends an earlier proposal. Like a proposal, the message involves both a commitment and a request. However, while a proposal is associated with a new task, an amendment is a suggested modification of an already-proposed task<br>*Refuse*: A refuse message rejects a meeting/action/task or declines an invitation/proposal<br>*Greet*: A greet message thank someone, congratulate, apologize, greet, or welcomes the recipient(s)<br>*Remind*: A reminder message reminds recipients of coming deadline(s) or threats to keep commitment |
| Response expectation [17]. | What type of response is expected? This dimension has ordinal values, consisting of:<br>*Ignore*: There is no realistic expectation that the recipients will properly read the email, let alone respond to them<br>*Accountable non-answer*: Recipient is expected to engage with the message or its attachments, but there is no reply required<br>*Postponed reply*: The messages requires a reply but not immediately<br>*Immediate reply*: The message requires a reply as soon as possible |
| Source authority [33] | 4-point ordinal scale (very low, low, high, very high): What is the authority of the sender? |
| Implicit reason | What was the reason to send the message? This categorization is based on the task-related categories in Enron, consisting of:<br>*Administrative procedure*: The message is part of an administrative procedure, such as financial arrangements or the organization of a meeting<br>*Legal procedure*: The message is part of a legal procedure<br>*Internal collaboration*: The message is part of a collaboration between people within the same company, such as messages related to internal projects<br>*External collaboration*: The message is part of a collaboration between people that are not working for the same company<br>*Travel planning*: The message is part of a travel plan, such as a confirmation of a hotel booking<br>*Employment arrangements*: The message is about employment arrangements, such as messages related to job seeking or job applications<br>*Logistic arrangements*: The message is about logistic arrangement. This includes general support and technical support<br>*Personal*: The message is of a personal, non work related, nature<br>*Other* |
| Number of tasks | How many tasks for the recipient are explicitly stated in the message (typically a number between 0 and 10)? |

A similar type of research has been done by Verberne et al. [33] and we will use the same approach. They developed a detailed scheme to assess the intent behind a query entered in a web search engine. Many of the dimensions they assess seem relevant for e-mail intent and task classification as well. More specifically, the action category which is based on the taxonomy by Broder [5]: informational, transactional and navigational, bares a strong resemblance to the categories of business communication: informative, requestive and directive [14], and to the e-mail acts defined by Carvalho and Cohen [6].

For that reason, we evaluate *e-mail act* as one of the dimensions in our classification scheme. Other dimensions from the literature that we will evaluate are *response expectation* [17] and *source authority* [33]. These are related to the message as a whole. On the message level we also evaluate the new dimensions *implicit reason* and *number of tasks*. The detailed description of the dimensions can be found in Table 1.[2]

For each task that is conveyed in a message we evaluate the following dimensions based on the query intent literature [33]: *spatial sensitivity*, *time sensitivity*, *task specificity* and *task topic*. Additionally we evaluate the new dimensions *task type* and *task subject*. A detailed description of these dimensions can be found in Table 2.[3]

### *3.2. Method*

In order to answer our research questions about reliability and validity of message intent dimensions we calculate the agreement between assessors. In our research we distinguish three types of assessors based on their relation to the e-mail message: (1) the assessor was the sender of the message, (2) the assessor was the recipient of the message, or (3) the assessor has no relation to the message (independent).

[2] Descriptions taken from Carvalho and Cohen [6].

[3] Retrieved from http://bailando.sims.berkeley.edu/enron_email.html

**Table 2**
Dimensions that describe a task that is to be undertaken by the recipient of the message and that was specified explicitly by the sender.

| Dimension | Description |
|---|---|
| Spatial sensitivity [33] | 4-point ordinal scale (very low, low, high, very high): Is the task associated with a certain location? For example is a meeting supposed to take place at a certain location, then the spatial sensitivity is very high; the task can only be executed there. |
| Time sensitivity [33] | 4-point ordinal scale (very low, low, high, very high): Is the task associated with a certain time? For example is the task supposed to be executed at a certain time, then the time sensitivity is very high. |
| Task specificity [33] | 4-point ordinal scale (very generic, somewhat generic, somewhat detailed, very detailed): How detailed is the description of the task? |
| Task type | What is the type of the task? categorical, consisting of:<br>*Physical*: The task requires physical action. For example ‘Do the groceries’ or ‘Get flowers’<br>*Informational*: The task requires knowledge. For example ‘When was Einstein born?’ or ‘Can you write a report about Einstein?’<br>*Procedural*: The task has a procedural nature; it is mainly administrative. For example ‘Can you plan a meeting’ |
| Task subject | What is the subject of the task/What is the task about? categorical, consisting of:<br>*Product*: e.g. ’Get flowers’<br>*Service*: e.g. ’Fix this problem for me<br>*Acknowledgment*: e.g. ’Write me a recommendation letter’<br>*Announcement*: e.g. ’Send a message that the meeting location has changed<br>*Decision*: e.g. ’Decide which flowers you prefer?’<br>*Reservation*: e.g. ’Confirm my reservation for room X’<br>*Event*: e.g. ’Make a schedule for event X’<br>*Meeting*: e.g. ’Confirm that you can meet at 10.30’<br>*Instructions*: e.g. ’Provide instructions how I can solve this bug’<br>*Collaboration*: e.g. ’Ask company X if they want to collaborate on topic Y’<br>*Information*: e.g. ’Provide the birth date of Einstein’<br>*Other* |
| Task topic [33] | Categorical, fixed set of categories from the well-known Open Directory Project (ODP), giving a general idea of what the topic of the task is. |

For this study, 5 collaborators have provided a total of 50 e-mail messages from their correspondence with the other collaborators. Each of them filled out a spreadsheet with columns corresponding to the various dimensions. The rows of the spreadsheet corresponded to the messages for which he or she was either the sender or the recipient. Furthermore, one independent individual (non-collaborator) who was not familiar with the context of the messages was asked to fill in the spreadsheet as well for all messages. All individuals were given the instructions for the dimensions as presented in Tables 1 and 2.

In this study we focus on two aspects of the annotations; *reliability* and *validity*, which we describe further in the next sections.

### *3.2.1. Reliability*

First we assess the agreement between assessors to determine the reliability of each dimension. Here we do not look at the relation of the assessor to the message (sender, recipient or independent). We calculate the inter-annotator reliability; how often do two annotators agree on their annotations for a dimension. The agreement on the dimensions was calculated using Cohen's $\kappa$ [9]. For the ordinal dimensions the agreement was calculated using weighted $\kappa$ [9]. The ordinal dimensions are: response expectation, source authority, number of tasks, spatial sensitivity, time sensitivity and specificity. All $\kappa$-agreements in this section are interpreted using the scale by Landis and Koch [24], where a $\kappa$ between 0.01–0.20 can be seen as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial and $>$ 0.80 as almost perfect agreement.

### *3.2.2. Validity*

Secondly, we assess the difference in agreement between sender–recipient assessor pairs and sender–independent assessor pairs to assess the validity of the dimension. Here we take the role of the assessor into account; was he the sender of the message, the recipient of the message or did he have no relation at all with the message. In fact our aim is to assess whether independent assessors can correctly interpret the original intent of the message. In this latter case we assume that the sender of the message knows his intent, so his annotation constitutes the ground truth.

To assess the validity of the assessment between sender–recipient annotator pairs and sender–independent pairs we use the pair-wise nature of the data (each message has been assessed by two pairs of annotators). We cannot use Cohen's $\kappa$ because it aggregates annotations over a complete dataset and cannot measure the agreement between two annotations for a single message. Therefore, we adopt the approach by Verberne et al. [33] where we compute per message a vector of scores for each of the assessor type pairs. For a given message, the annotation similarity between the two assessors of an annotator pair consists of Jaccard scores for categorical dimensions and normalized distances for the ordinal dimensions.[4]

[4] For the definition of the use of Jaccard in this case, we refer to [33].

**Table 3**
Agreement on the e-mail intent dimensions for sender–recipient (SR) pairs and sender–independent (SI) pairs, averaged over annotator pairs.

| Dimension | $\kappa$ SR | $\kappa$ SI |
|---|---|---|
| 1st e-mail act | 0.230* | 0.346* |
| 2nd e-mail act | 0.285* | 0.147 |
| Response expectation | 0.649* | 0.574* |
| Source authority | 0.263* | 0.160 |
| Implicit reason | 0.021* | 0.000 |
| Number of tasks | 0.664* | 0.556* |

*indicates significance of the $\kappa$ value at the 0.05 level.

Then we perform a pairwise significance test to compute the difference between annotation similarity by sender–recipient pairs and sender–independent pairs.

### 3.3. Results

We begin with an overview of the distribution of annotations. In this dataset, most messages contained a single task (55.6%), while 35.6% contained no task at all. There were no messages with more than 3 tasks. The main e-mail act was to deliver information (52.2%), followed by a request (21.7%). There was not often a necessity for immediate reply (6.5%): 37% required a postponed reply, while 56.6% required an accountable non-answer. The implicit reason for sending the message was mostly collaboration: 34.1% external collaboration and 43.2% internal collaboration.

More than half of the tasks were informational in nature (63.3%), while the remaining tasks were often procedural (30%). This is confirmed by the subject of the tasks that was often information (53.6%) or a decision (14.3%). Other common subjects of tasks were meeting, product or service (7.1% each).

#### 3.3.1. Reliability

To answer the question about which dimensions can be assessed reliably we look at the inter-annotator agreement. Dimensions where each annotator pair has at least a fair agreement are considered as reliable. In Table 3 we present the agreement between sender–recipient (SR) and sender–independent (SI) on the dimensions related to e-mail intent. We see a fair agreement on the first e-mail act, for both sender–recipient and sender–independent pairs. The agreement between sender and independent assessor on the second e-mail act was not significant, because there were too few annotations of the second e-mail act made by the independent assessor.

The agreement on response expectation is substantial for sender–recipient and moderate for sender–independent. This suggests that although an independent assessor can reliably estimate the response expectation, it is even easier for the recipient of a message.

In terms of source authority we see a fair agreement between sender and recipient, while the agreement between sender and independent assessor is slight and not significant. Since the agreement is low for the sender–independent pair (0.160) we decided to exclude this dimension from further experiments.

The implicit reasons in the message were assessed with only slight agreement between sender and recipient. The agreement between sender and independent assessor could not be calculated reliably as there was not enough variation in the annotations of the independent assessor compared to the annotations of the sender for the amount of data. On the basis of these agreements we should also remove the implicit reason dimension from further experiments. A detailed analysis reveals that the main reason for the low agreement is because of disagreement whether a message is considered to be external or internal collaboration. The distinction between the categories can be made by looking at the employer of the sender and comparing it to the employer of the recipient. Often this information can be extracted from the e-mail addresses. In this study, however, this information was not available, which made it difficult for the independent assessor to assess this dimension. We decided to keep the dimension in further experiments, but to make the e-mail addresses of sender and recipient part of the data.

The agreement on the number of tasks in the message is substantial between sender–recipient and moderate for sender–independent.

In Table 4 we present the agreement between sender–recipient and sender–independent on the dimensions related to the tasks in the e-mail messages.

There was a fair to moderate agreement on the spatial dimension. On the time sensitivity of tasks, the agreement between sender and recipient was much higher (substantial) than between sender and independent assessor (fair). This suggests that it is difficult for an independent assessor to reliably estimate the time sensitivity of a task. An explanation can be that the time assessment is made based on implicit information such as the past expectations between sender and recipient.

**Table 4**
Agreement on the task dimensions for sender–recipient (SR) pairs and sender–independent (SI) pairs.

| Dimension | $\kappa$ SR | $\kappa$ SI |
|---|---|---|
| Spatial sensitivity | 0.362* | 0.421* |
| Time sensitivity | 0.658* | 0.325* |
| Specificity | 0.230* | −0.211 |
| Type | 0.563* | 0.356* |
| Subject | 0.221* | 0.000 |
| Topic | −0.032 | −0.004 |

*indicates significance of the $\kappa$ value at the 0.05 level. This data was evaluated on 28 tasks.

The agreement on the specificity of the task was fair between sender and recipient, but negative between sender and independent assessor. Comments revealed that the assessors could not come to consensus about the interpretation of specificity, making this dimension hard to assess. Therefore this dimension was excluded in the remaining experiments.

The agreement on the type of the task was moderate between sender and recipient and fair between sender and independent assessor.

The agreement on the subject of the task could not be calculated between sender and independent assessor as there were too little data points for the number of categories in the dimension. Between sender and recipient the agreement was fair. On the basis of these results we have excluded the task subject dimension.

The agreement on the topic of the task was very low and not significant for both pairs of assessors. Since there was little variation in the general topic categories that could be assigned, this dimension was excluded in the remaining experiments.

### 3.3.2. Validity

To answer the question whether an independent assessor can assess the intent of a sender just as well as the recipient of a message, we assessed the difference in agreement between sender–recipient (SR) pairs and sender–independent (SI) pairs. This was calculated in a pair-wise fashion on message level as described in Section 3.2.2. The differences in agreement scores, significance values and effect size in terms of Cohen's *d* are reported in Table 5. From this we can conclude that an independent assessor is capable of interpreting the intent of the sender just as good as the recipient for the dimensions 1st e-mail act, response expectation, source authority and number of tasks ($P > 0.05$, so no significant difference between SI and SR).

However, the independent assessor is not able to interpret the implicit reason so well as the recipient. Therefore we should be careful in drawing conclusions based on this dimension.

The differences in agreement scores, significance values and effect size in terms of Cohen's *d* for the task dimensions are reported in Table 6. From this we can conclude that an independent assessor is capable of interpreting the tasks that the sender conveyed just as good as the recipient for all task dimensions. Nevertheless, caution should be taken when depending on the time dimension as the significance and effect size values indicate that it might be a dimension that is difficult to assess by an independent assessor.

**Table 5**
Difference in agreement between sender–recipient (SR) pairs and sender–independent (SI) pairs on message dimensions. Reported Jaccard scores are averaged over all messages.

| Dimension | Jaccard SR | Jaccard SI | *p*-value SR–SI | Cohen's *d* |
|---|---|---|---|---|
| 1st e-mail act | 0.52 | 0.63 | 0.23 | 0.22 |
| 2nd e-mail act | 0.61 | 0.20 | 0.00 | 0.93 |
| Response expectation | 0.76 | 0.67 | 0.36 | 0.19 |
| Implicit reason | 0.44 | 0.65 | 0.02 | 0.45 |
| Number of tasks | 0.92 | 0.91 | 0.64 | 0.09 |

**Table 6**
Difference in agreement between sender–recipient (SR) pairs and sender–independent (SI) pairs on task dimensions. Reported Jaccard scores are averaged over all messages.

| Dimension | Jaccard SR | Jaccard SI | *p*-value SR–SI | Cohen's *d* |
|---|---|---|---|---|
| Spatial sensitivity | 0.83 | 0.83 | 1.00 | 0.00 |
| Time sensitivity | 0.88 | 0.77 | 0.09 | 0.61 |
| Type | 0.77 | 0.73 | 0.72 | 0.05 |

**Table 7**
Inter-annotator agreement on the e-mail intent dimensions. All are significant at the 0.05 level.

| Dimension | $\kappa$ Enron | $\kappa$ Avocado |
|---|---|---|
| 1st e-mail act | 0.319 | 0.670 |
| Reply expectation | 0.328 | 0.585 |
| Implicit reason | 0.228 | 0.168 |
| Number of tasks | 0.334 | 0.667 |

## 4. E-mail intent assessments on larger datasets

In this second study we annotated part of a public dataset, Enron, and part of a licensed dataset, Avocado, of e-mail messages. We used the same classification scheme as described in Section 3.1 excluding the categories *source authority*, *task specificity*, *task subject* and *task topic*. These were excluded based on the results of the pilot study described in Section 3.

Details on the datasets can be found in Section 4.1. The aim of this study is to analyze task conveyance in e-mail message. Furthermore the annotated dataset will be shared with the research community to provide new opportunities for the development of (automated) e-mail support systems.

### *4.1. Data collection*

The data that we have annotated have been selected from the Enron and Avocado collections. The Enron dataset is a set of messages that was made public during a legal investigation of the Enron company [20]. It contains over 200,000 messages from 158 users that were sent or received between 1998 and 2004. The Enron company was an American energy, commodities and services company.

The Avocado collection[5] is a set of over 800,000 e-mail messages from the mailboxes of 279 users that were sent or received between 1995 and 2003. The data is collected from a defunct information technology company referred to as “Avocado”.

*Enron.* We selected a total of 1145 messages from the Enron dataset. Of these messages, 750 were randomly selected from the sent messages of the 15 most active users (50 each). The remaining 395 were coming from 15 randomly selected complete conversations. A conversation consists of all the messages sent between two individuals. These can contain multiple threads. Ten of the conversations were between two individuals within Enron, while 5 conversations were between an Enron-employee and an outsider. The average length of the selected internal conversations was 33.1 messages (minimum 6, maximum 81 messages). The external conversations had an average length of 14.4 (minimum 3, maximum 41 messages).

Each selected message was annotated according to the scheme in Section 3.1 using Amazon Mechanical Turk. Each message was annotated by 2 workers in order to make it possible to assess the inter-rater agreement. The annotators were required to have an annotation acceptance rate of more than 95% to ensure quality. A total of 3 messages were excluded from the final dataset because of noisy annotations, resulting in a dataset of 1143 annotated e-mail messages (see Table 7).

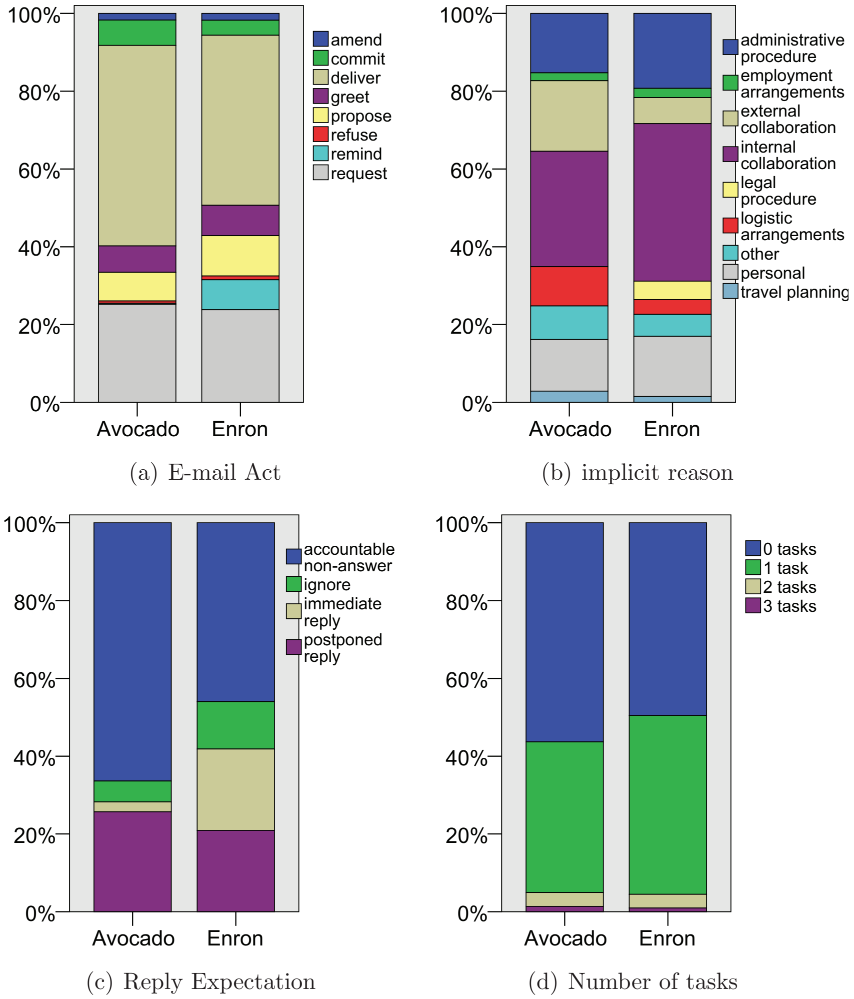
*Avocado.* A total of 379 messages was selected from the Avocado dataset. Of these messages, 250 were randomly selected from the sent messages of the 5 most active users (50 each) of which 7 messages were excluded because they were duplicates. The remaining 136 messages originated from 5 randomly selected complete conversations. A conversation consists of all the messages sent between two individuals. These can contain multiple threads. Of these conversations, 3 were between employees of Avocado, and 2 were between an Avocado-employee and an outsider. The average length of the selected internal conversations was 35 messages (minimum 10, maximum 71 messages). The external conversations had an average length of 16 (minimum 13, maximum 19 messages).

Each selected message was annotated according to the scheme in Section 3.1. Because of the license agreement, this set could not be annotated using Amazon Mechanical Turk. Instead, the data was annotated by two trained annotators (previously involved in the pilot study). They discussed the annotation dimensions in detail prior to annotating. All 379 messages were annotated by these trained annotators to assess agreement.

### *4.2. Results*

In this section we describe the analysis of the results of the annotations. We focus on assessing the agreement and present frequency distributions for the dimensions of interest. Additionally we analyze a conversation by means of a transition graph. A transition graph represents the transition probabilities of annotations within a conversation. For example, that a message annotated with “request” was followed by a message annotated with “deliver”

[5] https://catalog.ldc.upenn.edu/LDC2015T03



(a) E-mail Act

(b) implicit reason

(c) Reply Expectation

(d) Number of tasks

**Fig. 1.** Distribution of categorical values in e-mail dimensions.

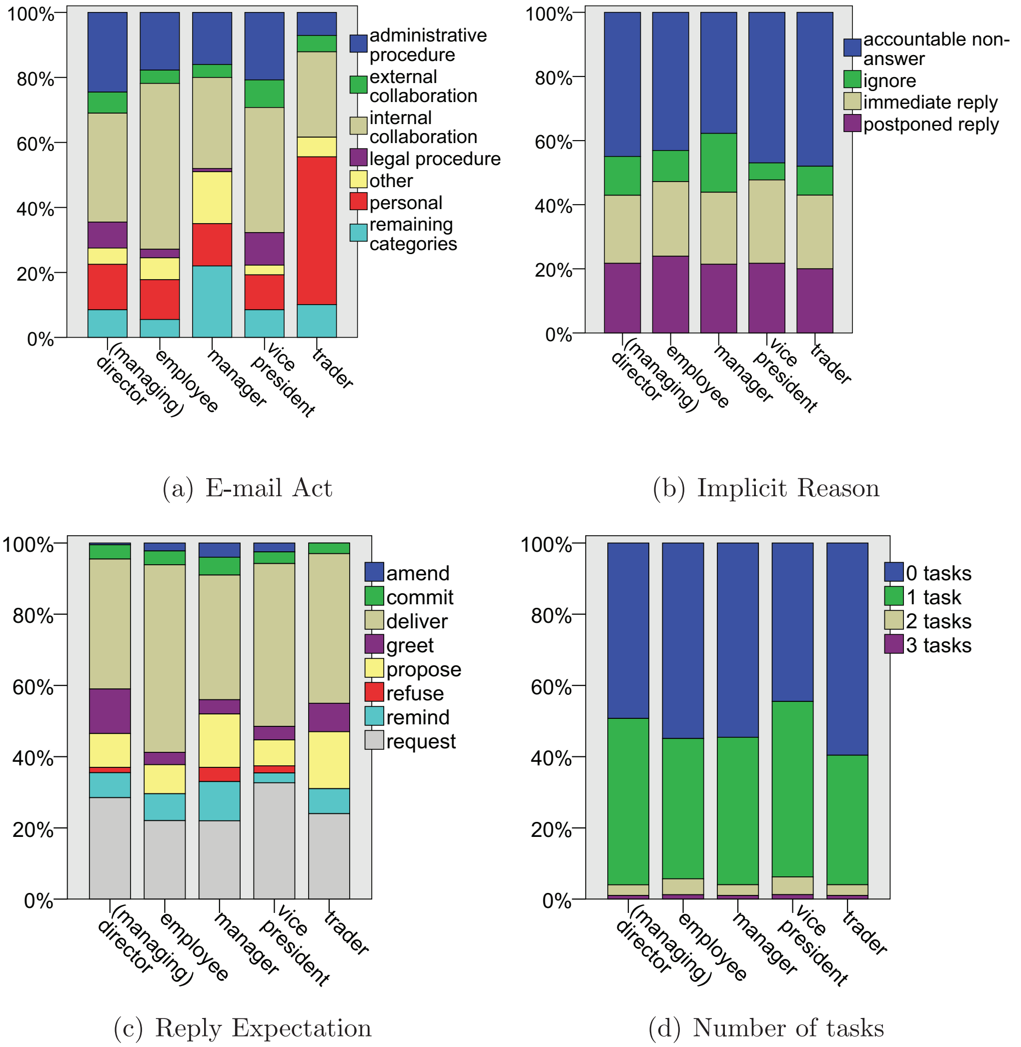
### 4.2.1. *E-mail intent*

We analyze the annotations on the message level. These dimensions are related to the e-mail message as a whole.

When we look at the agreement on the e-mail dimensions, the results show that for the Enron set the agreements are all fair. For the Avocado set the agreement is higher, being moderate or even substantial for all dimensions except the implicit reason, which has fair agreement. An analysis of the annotations shows that this dimension has a lower agreement because of confusion between the *logistic arrangements* category and the *internal collaboration* category as well as the *internal collaboration* and *external collaboration* category. The main activities in the Avocado company seem to be of a supportive and programmatic nature. Therefore many messages can actually be seen as both logistic as well as collaboration. Moreover, it is sometimes difficult to determine whether collaboration is within the company or externally. We have not reported the agreement on the 2nd e-mail act dimensions, as there were insufficient assessments made on the Enron dataset to assess it properly.

The high agreement on the Avocado set suggests that trained annotators reach higher agreement than non-trained annotators. Another explanation for the high agreement is that the messages in the Avocado set are easier to categorize. For both datasets the agreement is high enough to establish that the categorization can be assessed with at least a fair reliability. We cannot assess the validity of the assessments as the original senders and recipients are not available as assessors.

Fig. 1 shows the distribution of the annotation of the e-mail dimensions for the Enron and Avocado datasets. The distributions are averages over all annotations, by all assessors. We see that the distributions of the e-mail acts and reason for sending the message are very similar between the datasets. The e-mail acts *amend* and *refuse* are hardly ever used as main act in the message. There are also a few implicit reason categories that are not used very often: *travel planning* and *other*. In the Avocado set, *legal procedures* do not occur, where the Enron set contains a couple of messages related to legal procedures. This can be explained by legal issues that were surrounding Enron specifically.



(a) E-mail Act

(b) Implicit Reason

(c) Reply Expectation

(d) Number of tasks

**Fig. 2.** Distribution of categorical values in e-mail dimensions per role in the Enron dataset.

The reply expectation reveals that Enron messages are often a bit more urgent than Avocado messages (21% immediate reply vs. 7% immediate reply). Avocado messages are read without a reply in 67% of the cases whereas this is only 46% in Enron. When we look at task conveyance, the Avocado messages contain explicit tasks less often than the Enron messages (44% vs. 51%). Overall half of the messages do not contain a task. If the message does contain a task, it contains typically no more than one task.

For Enron we have information available about the various roles of the senders. We selected data from 2 directors, 5 employees, 1 manager, 1 trader and 4 vice presidents. These results are presented in Fig. 2. Here we can see that there are only slight variations in response expectation and number of explicit tasks based on employee role. We do see that normal employees seem to have a higher number of internal collaboration based messages, whereas managers send more employment arrangement related messages. Directors and vice presidents engage in more administrative procedures than normal employees and traders. Note, however, that as the data of the manager is only from one individual, these results are not generalizable. Overall the findings are not surprising and seem to comply with intuitions about office work.

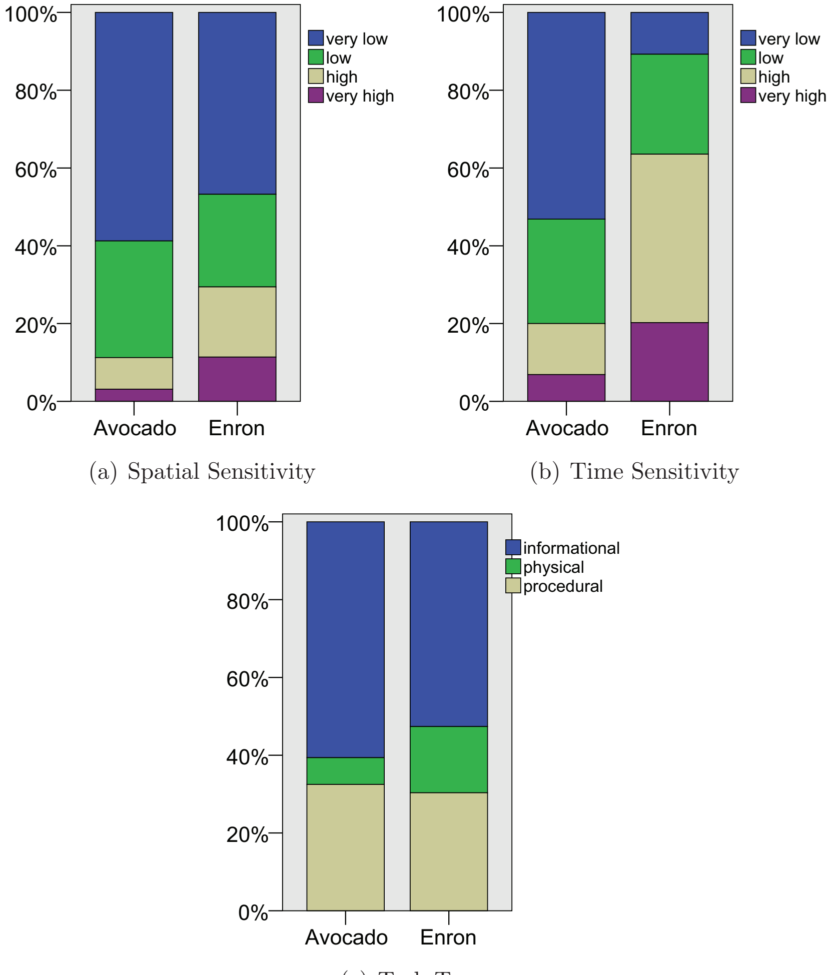
### *4.2.2. Task dimensions*

In this section we start again with the assessment of the inter-annotator agreement. We follow with an analysis of the distributions of annotations.

For the task dimensions we see slight to fair agreements on the Enron set, and fair to moderate agreements on the Avocado set (Table 8). This suggests that the task dimensions are harder to assess reliably, but that this can be improved with training (discussing the dimensions before annotation, like the expert annotators did). These agreements are calculated

**Table 8**
Inter-annotator agreement on the task dimensions. These results are for the most prominent task in a message as there are few messages with more than 1 task. All are significant at the 0.05 level.

| Dimension | $\kappa$ Enron | $\kappa$ Avocado |
|---|---|---|
| Spatial sensitivity | 0.187 | 0.344 |
| Time sensitivity | 0.138 | 0.449 |
| Type | 0.180 | 0.383 |



(a) Spatial Sensitivity

(b) Time Sensitivity

(c) Task Type

**Fig. 3.** Distribution of categorical values in task dimensions.

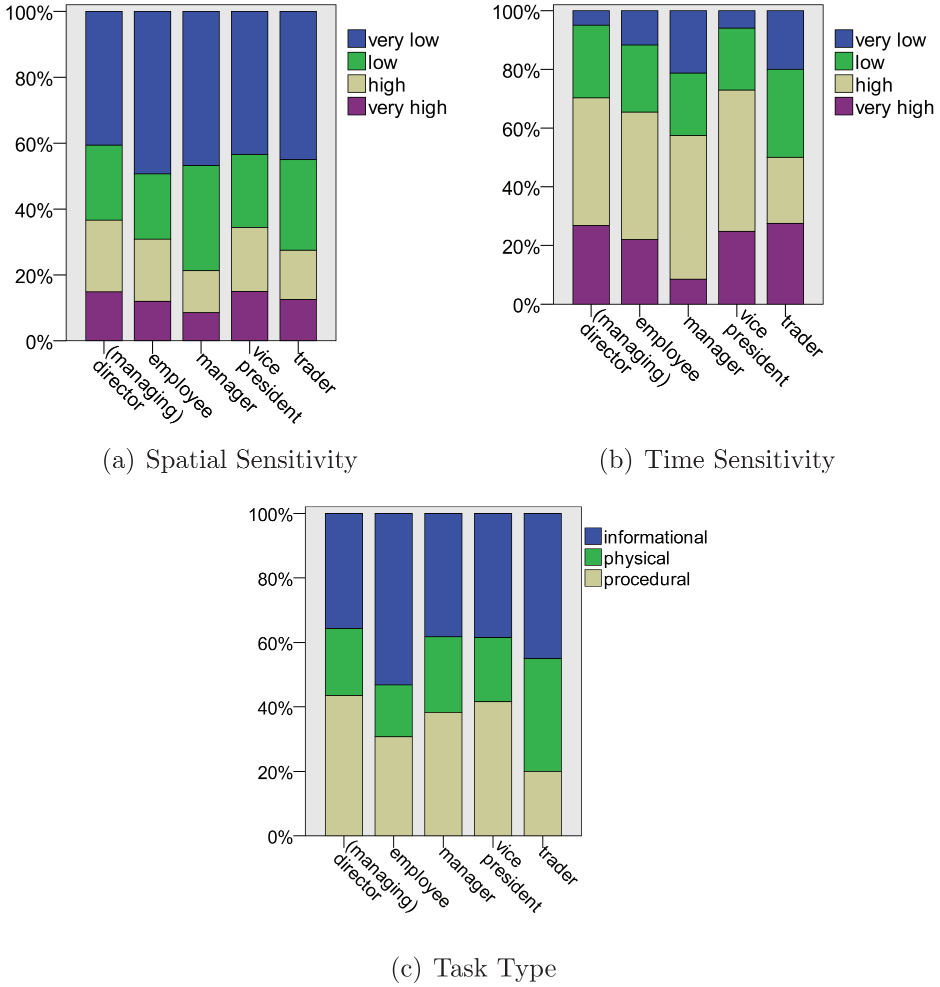
over the number of messages that contain at least one task, which are 379 messages for Enron and 139 messages for Avocado.

The distribution of the annotations of the task dimensions are presented in Fig. 3. The distributions are averages over all annotations, by all assessors. The distribution shows that the tasks in the Avocado set are typically less spatial and time sensitive than the tasks in Enron messages. The Enron messages contain physical tasks more often, whereas the percentage of procedural tasks in the Enron and Avocado messages are almost equal.

When we look at the task dimensions per role in the Enron dataset (Fig. 4) we see that employees send mostly informational tasks via e-mail. Vice presidents and managers send equally many procedural and informational tasks.

### *4.2.3. Evolution of a conversation*

In this section we look at how conversations evolve. We do this by analyzing e-mail acts and count how often they follow each other in a conversation, which we represent in a transition graph. An example of part of a conversation can be



(a) Spatial Sensitivity

(b) Time Sensitivity

(c) Task Type

**Fig. 4.** Distribution of categorical values in task dimensions per role in the Enron dataset.

**Table 9**
Part of a conversation from Enron.

| Sender | Recipient | Date | Subject | Body |
|---|---|---|---|---|
| Stan | Jim | 9-11-2001 13:55 | RE: Notre Dame | I would like you to get directly involved in selling the assets we have targeted for next year while overseeing the operations of Mariella, Pete and Orlando. That is a lot of stuff. However, I will keep you in mind as we figure out who is going to be on the transition team. |
| Jim | Stan | 19-11-2001 10:29 | FW: Draft MOU regarding equity sale | Wade/Rob/Bruce: I think we should reconfigure the MOU to constitute a binding obligation to purchase and sell rather than a MOU that would lead to a definitive agreement. We are trying to force an answer, it seems to me at this stage of the game, a definitive offer should be put forward for acceptance or rejection. Thoughts? Jim |
| Jim | Stan | 19-11-2001 15:06 | FW: ASSET SALES MEETING WITH STAN HORTON | Stan: Should not Ray Bowen or Jeff McMahon participate? Jim |
| Stan | Jim | 20-11-2001 05:28 | RE: ASSET SALES MEETING WITH STAN HORTON | I thought we should review the process with Mark and Jeff first. I will go ahead and invite either Jeff or Ray. |

found in Table 9. It is important to notice that a conversation includes all e-mail messages between two individuals. This is different from the so-called threads that are used in e-mail clients. A thread consists of all the messages between two individuals where the subject is the same, or with a prefix such as RE: or FWD:. A conversation can span multiple threads. In Fig. 5 we present the transition in e-mail acts for the conversations in the Enron and Avocado datasets.

For Enron it is interesting to note that a *request* message is often followed by a *deliver* message, a *commit* message or a *remind* message (Fig. 5(a)). This seems logical as the recipient can either full fill the request, commit to doing so at another time. When that does not happen the sender can remind him about the request. Furthermore it is interesting to see that

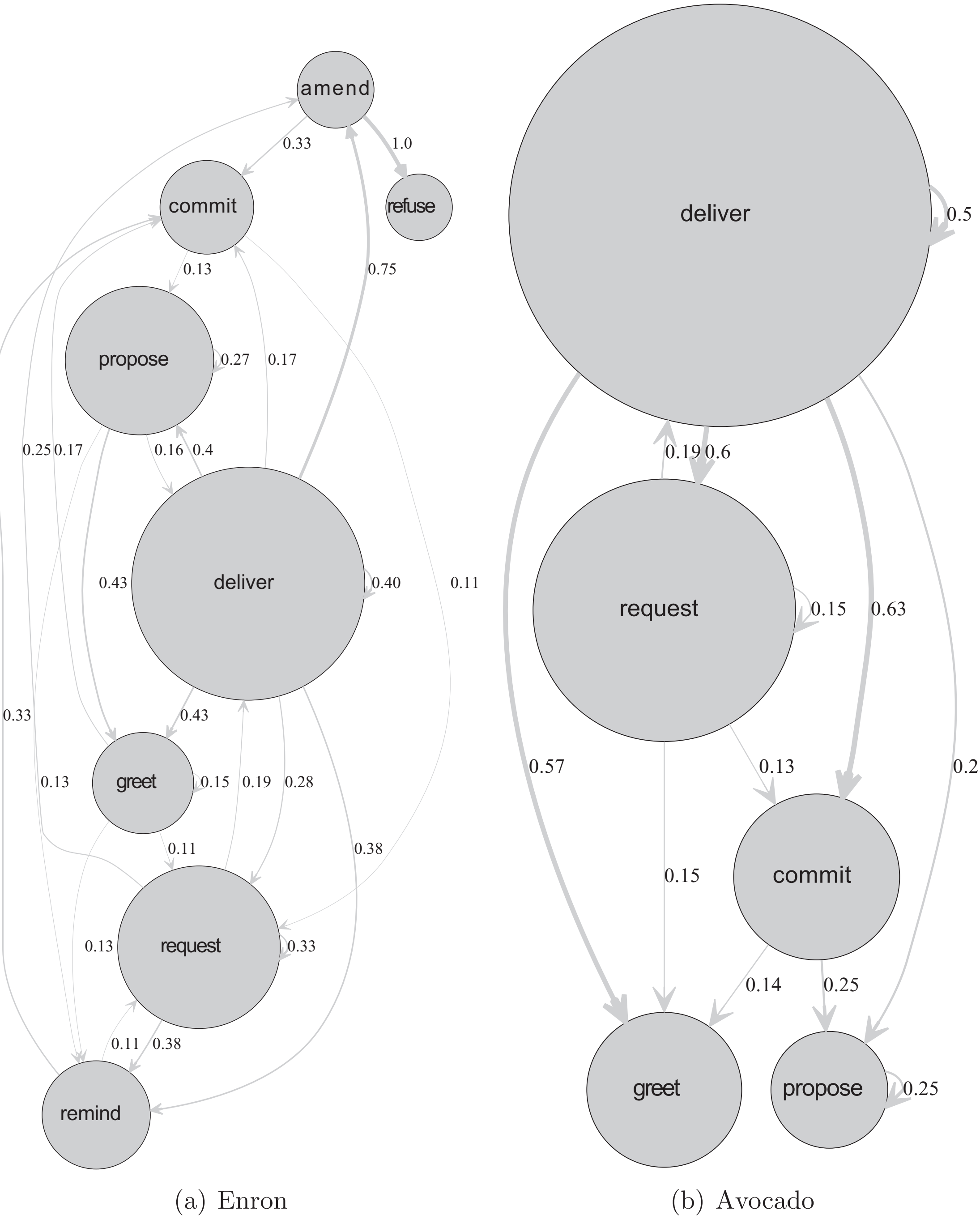


**Fig. 5.** Transitions in e-mail acts. Size of the ellipse represents the frequency of occurrence, the numbers represents the transition probabilities.

a *deliver* message or a *propose* message are often followed by a *greet*-type message. This could for example be a thank-you message. Another noteworthy point is that a refusal of a request seems to only occur after a request has been amended.

Similar trends seem to be going on in the Avocado messages: *requests* are mostly delivered or being committed to, and *deliver* messages are often followed by *greet* messages (most likely a thank you message). In contrast to Enron *greet* messages also follow after *request* and *commit* messages. Some e-mail act annotations, such as *refuse*, did not occur in the selection from the Avocado dataset.

Another interesting aspect to note in both Enron and Avocado is that there is a high probability that a *deliver* message is followed after another *deliver* message. This suggests that much information is delivered, even without a request. This can be for example because of own initiative, an earlier commitment, or because of agreements made outside the e-mail communication.

### 4.3. Discussion and limitations

In this section we have described the annotation, in terms of message intent and task conveyance, of two e-mail datasets that will be made available to the research community. We started with a pilot study in which we assessed the reliability and the validity of the various dimensions that we took under consideration to describe e-mail messages. This pilot study has resulted in a taxonomy for intent-based and task-based e-mail classification that consists of the following dimensions on message level (e-mail intent): *E-mail Act*, *Implicit Reason*, *Reply Expectation* and *Number of Tasks*. It consists of the following dimensions on the task level: *spatial sensitivity*, *time sensitivity* and *task type*.

The limitation of the reliability and validity research, as we have presented it in the pilot study, is that it was assessed on only a small dataset of 50 e-mail messages. The reason is that e-mail messages for research are hard to obtain because of the privacy concerns involved. To assess the reliability and validity we also need access to both the sender, the recipient and independent assessors, which makes it even harder to find useful messages. Finally, the task of annotating e-mail messages is labor intensive.

Despite the small dataset in the pilot study, we were able to obtain significant results. From these we can conclude that it is possible to assess the intent of a message and the tasks that were conveyed by independent assessors on all the dimensions in the taxonomy except for *implicit reason*. This suggests that the assessment of e-mail intent and tasks in e-mail messages is easier than query intent as there were few valid dimensions found by Verberne et al. [33] for query intent. This can probably explained by the amount of information available to the assessors: an e-mail message contains much more textual content than a query to base the assessment on.

In the main study, we annotated selections of the Enron and Avocado datasets using the task-based taxonomy. The limitation here is that we could not assess the validity as the original senders of the messages were not available. Furthermore, because of the strict license agreements for the Avocado dataset we could not use crowd-sourcing for annotation. This forced us to only annotate a very small portion of the messages (less than 1% of the messages). Nevertheless, the similarities between the distributions of the two datasets do not give reasons to doubt the representativeness of the data. In future works, a similar experiment could be repeated on different selections of Enron and/or Avocado data to obtain more annotated data.

A final limitation is in the analysis of the conversations. The conversations that were selected may not have been complete. Senders and recipients may have deleted messages before they were collected in the Enron or Avocado datasets. This may have distorted the analysis of the conversations.

These limitations show the challenges in working with e-mail data. E-mail messages are very sensitive to privacy concerns. Moreover the datasets are often incomplete because of messages that are deleted. Still, the annotations are an important contribution to the field. More effort should be taken to develop good annotated datasets of e-mail messages that can be shared with the research community, in order to make it possible to compare results.

## 5. Automated classification of e-mail intent

In this section we present an initial experiment for automated classification of e-mail intent, to investigate the feasibility of automatic classification in a future e-mail management application. A full automated classification study on all dimensions is out of the scope of this paper. Instead we focus on one of the dimension: *number of tasks*. This dimension could be used directly to indicate to a user whether an e-mail contains a task. We present a comparison of the performance of several well-known classification algorithms. Additionally we provide results on a cross-dataset study, where we use either Avocado or Enron to train the algorithms, and test on the other dataset.

### 5.1. Method

The data was collected from the annotated Avocado and Enron datasets. Only items on which the annotators agreed on the value of the dimension *number of tasks* were taken into account. This resulted in a dataset of 708 items for Enron, and 305 items for Avocado. The following features were extracted from the e-mail messages: tf∗idf weighted terms from subject, tf∗idf weighted terms (single words) from body, length of body (number of characters), number of sentences in body, number of question marks in body, number of recipients, number of sent messages by a sender

The following classifiers were used: a majority class classifier (ZeroR), k-NN with k = 5, SVM with a linear kernel and Naive Bayes, random forests and decision tree. Each algorithm was initialized with the default settings of scikit-learn [25]. For k-nn, SVM and Naive Bayes the tf∗idf message-body features were reduced to 50 components using LSA [16]. For random forest and decision Tree the number of term-features from the message-body was reduced by selecting those term-features that occurred more than 10 times in the data in order to optimize the results.

The results represent two experiments. In the first, single-set evaluation, we applied 10-fold cross validation. In the second, cross-set evaluation, we use one dataset (either Avocado or Enron) as training data and the other dataset for testing.

**Table 10**
Single-set classification performance, average accuracy (standard deviation) over 10-fold cross-validation.

| Classifier | Avocado | Enron |
|---|---|---|
| ZeroR (baseline) | 0.65 (+/− 0.03) | 0.52 (+/− 0.01) |
| k-NN (k=5) | 0.62 (+/− 0.14) | 0.50 (+/− 0.09) |
| SVM with linear kernel | 0.64 (+/− 0.06) | 0.53 (+/− 0.13) |
| Naive Bayes | 0.70 (+/− 0.12) | 0.67 (+/− 0.13) |
| Random forest | **0.71 (+/− 0.17)** | **0.72 (+/− 0.11)** |
| Decision tree | 0.70 (+/− 0.16) | 0.65 (+/− 0.17) |

**Table 11**
Cross-set classification accuracy.

| Classifier | Trained on Avocado, tested on Enron | Trained on Enron, tested on Avocado |
|---|---|---|
| ZeroR (baseline) | 0.52 | **0.63** |
| k-NN (k=5) | 0.51 | 0.45 |
| SVM with linear kernel | 0.53 | 0.44 |
| Naive Bayes | 0.62 | **0.63** |
| Random forest | 0.59 | 0.61 |
| Decision tree | **0.68** | 0.60 |

### *5.2. Results*

Table 10 shows that the baseline that is determined by a simple ZeroR classifier that guesses the majority class (0 tasks) from the data is quite high (65% in Avocado, 52% in Enron). In the Avocado data we can improve on this baseline with 6 percent point, where the Random Forest classifier yields the highest performance (71%), but is closely followed by Naive Bayes and decision tree (70%). For the Enron dataset we can improve on the ZeroR classifier with 20 percent point, where the random forest classifier again yields the highest accuracy.

For cross-set evaluation we see in Table 11 that training on Enron and testing on Avocado is hard since the Zero-R classifier already performs quite well (63%). When we train on the Avocado set to evaluate the Enron set, we can get an improvement over the baseline of 16 percent point with the decision tree classifier.

### *5.3. Discussion*

This experiment on automatically classifying the number of tasks in a message shows that it is not an easy task. First of all, the data was highly skewed. Most messages contained either 1 task, or no tasks at all. Only a few messages contained more tasks. This is a challenge for classifiers.

Additionally, it is important to find the right set of features for each dimension. As our cross-set evaluation showed, the necessary features can also depend on the classifier that is being evaluated. Nevertheless, cross-set evaluation is possible depending on the quality of the training data. Interestingly, the larger dataset (Enron) proved not suitable for cross-set evaluation. It is possible that tasks are less clearly formulated in the Enron dataset. This is supported by the finding that the inter-annotator agreement for the dimension number of tasks was significantly lower in Enron than in Avocado (Section 4.2.1)

For this experiment we have used default settings of the classifiers. Even then, we see that we can automatically classify the messages with reasonable accuracy. However, it is important to investigate the influence of the classifier parameters further. It is likely that with more extensive feature engineering and parameter optimization the accuracy can be improved further.

In future work, all dimensions should be explored for their potential in automated classification. Moreover, it is possible that each dimension has a different optimal feature set and optimal classifier, which should be explored. Finally, it should be investigated how to combine the various dimensions for optimal knowledge worker support.

## 6. Conclusion and future work

In this paper we presented a new taxonomy for an intent-based and task-based classification of e-mail messages. This taxonomy consists of the dimensions *e-mail act*, *reply expectation* and *number of tasks* that are assessed at the message level. The dimension *implicit reason* is assessed at the message level as well, but its reliability and validity should be investigated further. The task dimensions in the taxonomy consist of the dimensions *spatial sensitivity*, *time sensitivity* and *task type*. These are assessed for each task that is identified in an e-mail message.

The taxonomy was used in an annotation experiment with a selection of messages from the Enron and Avocado e-mail datasets. The resulting annotated datasets are available for future research.[6]

Finally, we presented a number of analyzes of the annotated data. These allow us to answer our research questions. The first research question was "To what extent do corporate e-mail messages contain tasks ?" We can conclude that approximately half of the e-mail messages contain a task. Moreover, typically only one task at a time is conveyed in a message. Furthermore, most messages are sent to deliver information or to request information. Requests are not often rejected. The implicit reason for sending messages is typically because of general collaboration, an administrative procedure or personal reasons. In terms of reply expectation, about half of the messages do not require a reply. If a reply is needed, it typically does not need to be immediate.

For our second research question "What are the characteristics of tasks in e-mail messages ?" We can conclude that most tasks can be executed everywhere (low spatial sensitivity). Some tasks do have a high or very high time sensitivity such as a deadline, but the likeliness of this happening depends strongly on the company. The type of the task is mostly informational or procedural. This corresponds to the work contexts in which both e-mail datasets have been collected: a knowledge worker environment where the exchange of information is an important part of the work.

About the third research question "How does a work-related e-mail conversation evolve ?" We can conclude that there is a high probability that a *deliver* message is followed after another *deliver* message. This suggests that much information is delivered, even without a request. Furthermore, *requests* are mostly delivered or being committed to, and *deliver* messages are often followed by *greet* messages. These *greet* messages are most likely thank you messages.

The annotations on these datasets can be used for research into (automatic) task-based categorizations of messages. We have done so in order to answer our fourth and final research question "Can we determine the number of tasks in a message automatically ?". We can conclude that this is possible with 71% accuracy. Most likely this number can improve with more extensive feature engineering and classifier parameter optimization. Moreover we showed that it is also possible to use the data from the Avocado set to classify the Enron dataset. This is achieved with a decision tree classifier, which suggests that it is possible to find common rules about the number of tasks in a message. Note, however, that Enron and Avocado originate from businesses that share certain qualities. Therefore, in order to substantiate this result, research on a more diverse set of data is required.

In general, detecting how many tasks a message contains, whether a reply is expected, or what the spatial and time sensitivity of such a task is, can help in providing a more detailed priority-estimation of the message for the recipient compared to existing work [1,29]. Such a priority-based categorization can support knowledge workers in their battle against e-mail overload. For this reason, future work should be directed at a more thorough exploration of automatic classification on the dimensions in the taxonomy. This requires research to which dimensions can be assessed using machine learning techniques, which features optimally model each dimension and which classifiers are best suited for the task.

## Acknowledgments

This publication was supported by the Dutch National Program COMMIT (Project P7 SWELL).

## References

[1] D. Aberdeen, O. Pacovsky, A. Slater, The learning behind gmail priority inbox, in: Proceedings of the NIPS 2010 Workshop on Learning on Cores, Clusters and Clouds (LCCC), 2010.

[2] D. Bawden, L. Robinson, The dark side of information: overload, anxiety and other paradoxes and pathologies, J. Inf. Sci. 35 (2) (2009) 180–191.

[3] R. Bekkerman, Automatic categorization of email into folders: Benchmark experiments on Enron and SRI corpora, Comput. Sci. Dep. Fac. Publ. Ser. (2004) 218.

[4] J.C. Benselin, G. Ragsdell, Information overload: The differences that age makes, J. Librariansh. Inf. Sci. (2015).

[5] A. Broder, A taxonomy of web search, in: Proceedings of the 2002 ACM SIGIR Forum, vol.36, ACM, 2002, pp. 3–10.

[6] V.R. Carvalho, W.W. Cohen, On the collective classification of email speech acts, in: Proceedings of the Twenty-eighth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2005, pp. 345–352.

[7] M. Cecchinato, A.L. Cox, J. Bird, I check my email on the toilet: Email practices and work-home boundary management, in: Proceedings of the 2014 MobileHCI Workshop, 2014.

[8] S. Chakravarthy, A. Venkatachalam, A. Telang, A graph-based approach for multi-folder email classification, in: Proceedings of the 2010 IEEE Tenth International Conference on Data Mining (ICDM), IEEE, 2010, pp. 78–87.

[9] W.W. Cohen, V.R. Carvalho, T.M. Mitchell, Learning to classify email into "speech acts", in: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2004, pp. 309–316.

[10] M. Dredze, T. Lau, N. Kushmerick, Automatically classifying emails into activities, in: Proceedings of the Eleventh International Conference on Intelligent User Interfaces, ACM, 2006, pp. 70–77.

[11] M. Dredze, T. Brooks, J. Carroll, J. Magarick, J. Blitzer, F. Pereira, Intelligent email: Reply and attachment prediction, in: Proceedings of the Thirteenth International Conference on Intelligent User Interfaces, ACM, 2008, pp. 321–324.

[12] J. Gains, Electronic mail a new style of communication or just a new medium?: An investigation into the text features of e-mail, Engl. Specif. Purp. 18 (1) (1999) 81–101.

[13] J. Gantz, A. Boyd, S. Dowling, Cutting the clutter: Tackling information overload at the source, International Data Corporation White Paper, 2009.

[14] M. Ghadessy, J. Webster, Form and function in English business letters: Implications for computer-based learning, Registers of Written English: Situational Factors and Linguistic Features, 1988, pp. 110–127.

[15] M. Grbovic, G. Halawi, Z. Karnin, Y. Maarek, How many folders do you really need?: Classifying email into a handful of categories, in: Proceedings of the Twenty-third ACM International Conference on Conference on Information and Knowledge Management, ACM, 2014, pp. 869–878.

[6] http://cs.ru.nl/~msappelli/data/

[16] N. Halko, P.-G. Martinsson, J.A. Tropp, Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions, SIAM Rev. 53 (2) (2011) 217–288.

[17] B.V. Hanrahan, M.A. Pérez-Quiñones, D. Martin, Attending to email, Interact. Comput. (2014).

[18] A. Kalia, H.R. Motahari Nezhad, C. Bartolini, M. Singh, Identifying Business Tasks and Commitments from Email and Chat Conversations, Technical Report, 2013. Technical Report, HP Labs.

[19] Y.M. Kalman, G. Ravid, Filing, piling, and everything in between: The dynamics of e-mail inbox management, J. Assoc. Inf. Sci. Technol. (2015).

[20] B. Klimt, Y. Yang, Introducing the Enron corpus., in: Proceedings of the First Conference on Email and Anti-Spam (CEAS), 2004.

[21] F. Kooti, L.M. Aiello, M. Grbovic, K. Lerman, A. Mantrach, Evolution of conversations in the age of email overload, in: Proceedings of the Twenty-fourth International Conference on World Wide Web, 2015.

[22] Y. Koren, E. Liberty, Y. Maarek, R. Sandler, Automatically tagging email by leveraging other users' folders, in: Proceedings of the Seventeenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2011, pp. 913–921.

[23] A. Lampert, R. Dale, C. Paris, Requests and commitments in email are more complex than you think: Eight reasons to be cautious, in: Proceedings of the 2008 on Australasian Language Technology Association, vol. 6, 2008, pp. 64–72.

[24] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics (1977) 159–174.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[26] K. Peterson, M. Hohensee, F. Xia, Email formality in the workplace: A case study on the Enron corpus, in: Proceedings of the 2011 Workshop on Languages in Social Media, Association for Computational Linguistics, 2011, pp. 86–95.

[27] F.P. Romero, J.A. Olivas, P.J. Garcés, Fzmail: Using FIS-CRM for e-mail classification., JACIII 11 (1) (2007) 40–50.

[28] M. Sahami, S. Dumais, D. Heckerman, E. Horvitz, A Bayesian approach to filtering junk e-mail, in: Proceedings of the 1998 Workshop on Learning for Text Categorization, vol.62, 1998, pp. 98–105.

[29] M. Sappelli, S. Verberne, W. Kraaij, Combining textual and non-textual features for e-mail importance estimation, in: Proceedings of the Twenty-fifth Benelux Conference on Artificial Intelligence, 2013.

[30] M. Sappelli, S. Verberne, W. Kraaij, E-mail categorization using partially related training examples, in: Proceedings of the Fifth Information Interaction in Context Symposium, 2014.

[31] R.B. Segal, J.O. Kephart, Mailcat: An intelligent assistant for organizing e-mail, in: Proceedings of the Third Annual Conference on Autonomous Agents (AGENTS '99), ACM, New York, NY, USA, 1999, pp. 276–282, doi:10.1145/301136.301209.

[32] J.R. Tyler, J.C. Tang, When can i expect an email response? A study of rhythms in email usage, in: Proceedings of the Eighth Conference on European Conference on Computer Supported Cooperative Work (ECSCW), Springer, 2003, pp. 239–258.

[33] S. Verberne, M. Heijden, M. Hinne, M. Sappelli, S. Koldijk, E. Hoenkamp, W. Kraaij, Reliability and validity of query intent assessments, J. Am. Soc. Inf. Sci. Technol. 64 (11) (2013) 2224–2237.

[34] S. Whittaker, C. Sidner, Email overload: exploring personal information management of email, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 1996, pp. 276–283.