A Large-Scale Corpus of E-mail Conversations with Standard and Two-Level Dialogue Act Annotations

Motoki Taniguchi[†,‡], **Yoshihiro Ueda**[†], **Tomoki Taniguchi**[†] and **Tomoko Ohkuma**[†]
{motoki.taniguchi, ueda.yoshihiro, taniguchi.tomoki, ohkuma.tomoko}@fujixerox.co.jp
[†] Fuji Xerox Co., Ltd.
[‡] Tokyo Metropolitan University

Abstract

We present a large-scale corpus of e-mail conversations with domain-agnostic and two-level dialogue act (DA) annotations towards the goal of a better understanding of asynchronous conversations. We annotate over 6,000 messages and 35,000 sentences from more than 2,000 threads. For a domain-independent and application-independent DA annotations, we choose ISO standard 24617-2 as the annotation scheme. To assess the difficulty of DA recognition on our corpus, we evaluate several models, including a pre-trained contextual representation model, as our baselines. The experimental results show that BERT outperforms other neural network models, including previous state-of-the-art models, but falls short of a human performance. We also demonstrate that DA tags of two-level granularity enable a DA recognition model to learn efficiently by using multi-task learning. An evaluation of a model trained on our corpus against other domains of asynchronous conversation reveals the domain independence of our DA annotations.

1 Introduction

The recent growth of textual communication media such as e-mails and online forums has led to the great demand for techniques enabling the automatic analysis of conversational structures from texts for information retrieval and intelligent assistance. Dialogue acts (DAs), which are also known as "speech acts" in some studies, are one of such conversational structures. The DAs of text in conversations are defined as having communicative functions, such as asking questions, requesting some information, and offering suggestions. Analyzing the DAs in conversations helps many downstream applications, including summarization (Bhatia et al., 2014; Oya and Carenini, 2014), question answering (Hong and Davison, 2009), and conversational agents (Peskov et al., 2019).

Conversations fall into synchronous conversations (e.g., phone calls and meetings) and asynchronous conversations (e.g., e-mails and online forums). In a synchronous conversation all participants engage at the same time with all participants, whereas in an asynchronous conversation participants interact with each other at different times. As is well known, the asynchronous properties make the conversational flow of asynchronous conversations different from those of synchronous conversations (Joty et al., 2013; Louis and Cohen, 2015). Topics in asynchronous conversations are often interleaved and not contiguous sequence. This complexity in a conversational flow makes DA recognition in asynchronous conversations a challenging task compared to synchronous conversations, particularly when the thread structure (reply relations) is missing.

The existing corpora for DA recognition in asynchronous conversations have some of the shortcomings: (i) **Scale**: Whereas large-scale corpora such as the Meeting Recorder Dialog Act (MRDA) (Shriberg et al., 2004) and the Switchboard Discourse Annotation and Markup System of Labeling (SWBD) (Jurafsky et al., 1997) are available in synchronous conversation, most of the available corpora in asynchronous conversations are limited to a few thousands messages or sentences, and are insufficient to train more expressive models of DA recognition. (ii) **Annotating scheme**: Annotation schemes used in the existing corpora are designed for a particular purpose or a particular application domain. For instance,

Proceedings of the 28th International Conference on Computational Linguistics, pages 4969–4980
Barcelona, Spain (Online), December 8-13, 2020

Kim et al. (2010) focuses on technical help forums and defines domain-specific DA tags such as *RESOLUTION*, which represents a post confirming that an answer operates based on its implementtation. (iii) **Annotation granularity**: Many of the existing corpora in asynchronous conversations are annotated with DA tags at only either sentence-level or message-level (comment-level). However, the level of DA tags required for a conversational analysis depends on its applications. For instance, sentence-level DAs can help with a summarization of a message, and message-level DAs are beneficial for mining question-answer pairs from a community question-answering site. Therefore, multi-level annotations of DAs allow for wider range of applications. Moreover, multi-level annotations can provide a model of DA recognition to learn efficiently by multi task learning.

To fulfil all the requirements, we present a large-scale corpus of e-mail conversations with domain-agnostic and two-level DA annotations. As our source of asynchronous conversations, we use the Enron e-mail dataset (Klimt and Yang, 2004) because e-mail is one of the most used communication media and the number of publicly available raw texts from e-mails is larger than that of other datasets. Our corpus consists of over 2,000 threads, 6,000 messages, and 35,000 sentences. To conduct domain- and application-independent annotations, we adopt the ISO standard 24617-2 (Bunt et al., 2012) as our DA annotation scheme. The standard is designed for application- and domain-agnostic tags of DAs. As a segment unit of a DA, we adopt both a sentence and a message.

To assess the difficulty of DA recognition on our corpus, we evaluate several models, including a pre-trained contextual representation model, as our baselines. We also evaluate a human performance on our corpus for a model comparison. In addition, we demonstrate that DA tags of two-level granularity enable a DA recognition model to learn efficiently by applying multi-task learning. To evaluate the domain independence of our DA annotations, we investigate the domain adaptation, wherein a model is trained on our corpus as a source domain and is applied to other corpora as the target domains.

The contributions of this paper are as follows:

- We develop a large-scale corpus of e-mail conversations with domain-agnostic and two-level DA annotations that satisfy all the requirements of scale, annotation scheme, and annotation granularity.
- To evaluate the difficulty of DA recognition on our corpus, we demonstrate that BERT outperforms earlier approaches with neural sequence learning but falls short of a human performance of DA recognition on our corpus .
- Empirical results show that a model trained on our corpus can be easily applied to other domains in asynchronous conversations.

2 Related studies

2.1 Discourse Act Corpus

There are several large corpora of synchronous conversations annotated with speech acts. SWBD (Jurafsky et al., 1997) comprises 205,000 utterances of one-on-one phone conversations from the Switchboard corpus (Godfrey et al., 1992). This corpus is annotated with a discourse tag-set extended from the Discourse Annotation and Markup System of Labeling (DAMSL) (Core and Allen, 1997). In addition, MRDA (Shriberg et al., 2004) contains over 180,000 in face-to-face meetings with DA tags defined in Dhillon et al. (2004).

There have also been several studies on annotated corpora with DA tags in asynchronous conversations. Cohen et al. (2004) developed a corpus that contains 1,357 e-mails and annotated DA tags according to an ontology of verbs and nouns. Ravi and Kim (2007) annotated 1834 messages (475 threads) of an online discussion site with speech act tags. Kim et al. (2010) collected 320 threads containing a total of 1,332 posts from newsgroup data to annotate with DAs. Bhatia et al. (2012) also annotated 556 posts of 100 threads in Ubuntu forum with speech act tags. Jeong et al. (2009) created two corpora for sentence-level speech act recognition. The corpora consist of 40 e-mail threads and 100 threads selected from an online travel forum. Feng et al. (2006) annotated 2214 messages in 640 threads of an online

forum with speech act tags. Joty and Hoque (2016) tagged 1565 sentences in 50 threads in a community question answering site with DA tags. BC3 (Ulrich et al., 2008) consists of 40 e-mail threads and 3222 sentences from mailing lists of the World Wide Web Consortium. The sentences is annotated with five speech act tags for summarization. All of the above corpora contains a few thousands of labeled messages or sentences at most and are smaller than our corpus. One exception that is larger than ours is the corpus of Zhang et al. (2017). It consists of 9, 483 threads with 115, 827 comments from a social news site. While the size of their corpus is large, the annotation scheme is specific for online forums and annotated at only comment-level.

2.2 Dialogue Act Recognition in Asynchronous Conversations

Cohen et al. (2004) built a supervised classifier with textual features for predicting message-level DA. Carvalho and Cohen (2005) extended the model to capture the sequential correlation among messages in the same thread by using a dependency-network based collective classification method. Kim et al. (2010) introduce structural learners including Conditional Random Fields (CRF) (Lafferty et al., 2001) to be optimized as a sequential labeling problem.

More recently, Joty and Hoque (2016) firstly introduced a neural network approach base on a combination of a long short term memory (LSTM) and a CRF layers. In their approach, the LSTM and CRF layers are trained separately. Joty and Mohiuddin (2018) demonstrated that the word embeddings that are pre-trained on conversations boosts the performance of a neural network based model for speech act recognition. The work of Joty and Hoque (2016) was extended to train the model in an end-to-end fashion and introduce hierarchical LSTM of sentence and conversation level (Mohiuddin et al., 2019). To the best of our knowledge, these model are the state of the art, hence we adapt these models as our baselines to analyze the characteristics of our corpus.

3 Dialogue Act Annotation

3.1 Dialogue Acts

We adopt the ISO standard 24617-2 (Bunt et al., 2012) as our DA annotation scheme. This standard has been developed owing to the need for a domain- and application-independent DA annotation scheme. Although a DA of the standard has several ingredients, we focus on communicative functions in this paper. Communicative functions are defined as a hierarchical taxonomy for general use. We restricted the taxonomy according to the frequency of the DA tags in a trial annotation of small data. The restricted version of the ISO annotation scheme is shown in Figure 1. We added *None* and *Others* to the tag set of the restricted taxonomy because of the noisy nature in the messages of the Enron dataset. As a segment unit of a DA, we adapt both a sentence and a message. Although a segment (message or sentence) could have multiple DA tags, we annotate the only tag that represents its main communicative function.

The definitions of each DA tag are as follows. ***Request*:** A segment of text that asks the recipient to perform a certain activity. ***Suggestion*:** A segment of text that contains an idea or plan mentioned by the sender. ***Commissive*:** A segment of text that commits the sender to certain actions in the future. ***Question*:** A segment of text that asks a question. ***Answer*:** A segment of text that answers a question. ***Inform*:** A segment of text that provides information to the recipient. ***None*:** A segment of text that is provided with less informative communication, such as a greeting or joke. ***Others*:** A segment of text that is provided as a part of a list, table, or signature.

3.2 Data Collection

We used the Enron email dataset (Klimt and Yang, 2004) as the source of our corpus. The Enron e-mail dataset is a collection of e-mails released for an investigation into the Enron corporation, and contains over six million messages belonging to 158 users.

The thread structure is missing in the Enron e-mail dataset. Hence, we create a thread by splitting a single original message by markers of reply or forward templates (e.g. "—–original message—–" and "—–forwarded by John Doe 01/23/2000 04:56 PM—–"). Duplicated threads containing the same messages were filtered out. We discarded the threads contains only one message or more than 20 messages

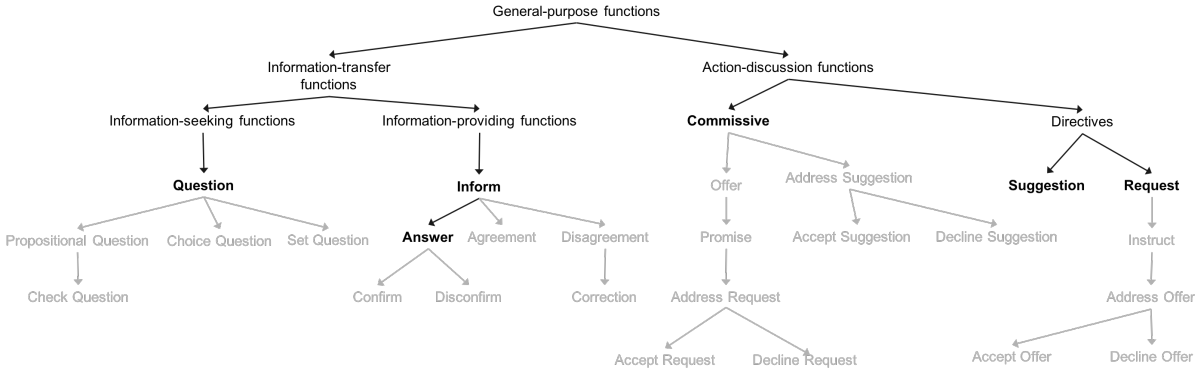


Figure 1: Communicative functions from ISO 24617-2 general purpose dimension. The grayed-out nodes of the taxonomy are not used, and bold nodes are applied as a tag set of DAs in our corpus.

MID	SID	Text	S-tag	M-tag
1	1	To: All Enron Employees Transferring to X Corp. (or its affiliates)	Inform	Request
	2	Under the terms of the various agreements among certain Enron entities and X certain Enron data and documents may be provided to X.	Inform	
	3	To ensure compliance with those agreements, Enron's document retention policy, as well as directives of various investigative agencies, the attached protocol regarding the process for migration of any information or data to X has been developed.	Inform	
	4	Please review the protocol carefully.	Request	
	5	Then, complete the attached Certification and return it, with attachments, to Bob no later than January 21, 2001.	Request	
		(Omitted)		
	14	Thank you for your assistance in making the transition to X a success.	Others	
2	1	OK	Answer	Question
	2	Who is the individual responsible for making sure that 600 plus people will read the 9 pages of protocol and will also verify and collect a signed certification from every new X Crop. employee by this Friday?	Question	
	3	Who will collect Enron's copy of these and hold them and is someone coordinating with X Corp. directly?	Question	
	4	Dave	None	
3	1	Carol, Correct me on this one if I am wrong...	Question	Answer
	2	In the body of the email distributed yesterday afternoon it says that the form should be sent to Bob.	Inform	
	3	I believe she is in the legal department for Enron.	Inform	
	4	So I would assume she is the person who will track and report on who has and has not submitted their certification letter.	Answer	
	5	Isaac	None	
4	1	Yes, you are correct.	Answer	Answer
	2	Carol	None	

Table 1: Example e-mail conversation and our DA annotations. MID and SID represent a message and sentence ID, respectively. S-tag and M-tag correspond to the sentence- and message-level tags.

due to the annotation efficiency. We used Sentencizer in spaCy [1] to split a message into sentences.

We found that the distribution of DA tags in the e-mail dataset is imbalanced. Owing to the imbalance, a random sample of threads does not contain a sufficient number of some tags. To improve the annotation efficiency, we collected threads by identifying messages containing cue phrases that indicate tags of low frequency. The cue phrases (e.g., those ending with "?" for *Question* and "let me know" for *Request*) are created manually based on careful observations. Eventually, we provided annotators with 2,672 threads sampled randomly from the collected threads.

3.3 Annotation

One of the practical approaches to create a large-scale corpus is to use crowdsourcing. However, crowd workers often provide noisy annotations. Hence, in addition to crowd workers, we employed experts as annotators to evaluate the annotation quality of the crowd workers and establish a human performance in terms of DA recognition. The expert annotators are well-trained for linguistic annotations. We held out 98 of the sampled threads to evaluate the annotation quality of crowd workers. Each thread of the held-out set was assigned to three crowd workers and three experts. By contrast, each of the remaining threads was annotated by a single crowd worker to improve the annotation scalability. The annotators were asked to assign a DA tag to each sentence and message of the thread, respectively. Table 1 shows an example of an e-mail conversation and our annotation of DA tags.

We use Fleiss's kappa κ (Fleiss, 1971) as a measure of inter-annotator agreement. The sentence-level

[1] https://spacy.io/

κ between the crowd workers is 0.56, and the message-level κ is 0.50, indicating moderate agreement. The sentence-level κ between the experts is 0.85, and the message-level κ is 0.71, indicating substantial agreement. Although the annotation quality of the crowd workers is not as high as that of the experts, the larger number of annotation data significantly improves the DA recognition models (see Section 6.4).

4 Dataset Analysis

4.1 Textual Statistics

Table 2 summarizes the statistics of our corpus and the existing corpora of asynchronous conversations BC3 and QC3. We use these existing corpora for comparison because they are publicly available and the annotation schemes are relatively similar to ours. BC3 (Ulrich et al., 2008) is also an e-mail corpus derived from mailing list threads. QC3 (Joty and Hoque, 2016) is made up of conversations of an online forum. Compared to the existing corpora, the number of threads in our corpus is larger by two orders of magnitude. Messages in our corpus are comparable in length to those in BC3 and longer than those in QC3 in terms of the average number of sentences in a message (5.17 vs. 5.24 and 2.50 sentences per message). If we arrange the sentences in the messages of a thread in chronological order, the distance of the adjacency pairs (such as a question and answer) (Schegloff and Sacks, 1973) across the message may be long in our corpus. Therefore, it is important for the modeling of the DA recognition in our corpus to capture such a long dependency of the DA tags.

	Ours	BC3	QC3
# of threads	2,574	39	47
# of messages	6,636	254	626
# of sentences	34,323	1,332	1,565
# of words/sentence	10.5	11.5	20.6

Table 2: Text statistics of corpora in asynchronous conversation.

Tag	Message	Sentence
Request	15.2	10.0
Suggestion	4.2	4.7
Commissive	3.0	2.8
Question	11.3	6.1
Answer	11.9	3.5
Inform	41.7	31.4
None	7.5	29.8
Others	5.3	11.6

Table 3: Distribution of DA tag (in percentage) at message-level and sentence-levels in our corpus.

4.2 Distribution of Dialogue Act Tag

Table 3 shows the distribution of DA tags in our corpus. *Inform* is the most frequent tag at either a sentence or message level, which agrees with our intuition. The second-most frequent tags are a *Request* at the message level and *None* at the sentence level. A *None* sentence occurs frequently, which is likely due to the fact that Enron Co. is an energy company and that their e-mails often contain a list or table of their products. The top two tags account for over 55% of the total. This suggests that the tag distribution in our corpus is imbalanced.

For a comparison with BC3 and QC3, we converted our DA scheme into their scheme of speech acts by using the mapping shown in Table 4. Because BC3 and QC3 are available only at the sentence-level speech act tags, we conducted a comparison at the sentence level. Figure 2 shows a comparison of the tag distributions among our corpus, BC3, and QC3. The entropy of the tag distribution in our corpus is higher than that of BC3 and QC3 (0.61 vs. 0.40 and 0.52). Therefore, we can regard our corpus as having more balanced distributions.

4.3 Dependencies Between Dialogue Act Tags

The importance of a sequential relationship between adjacent DAs (such as a question and answer) for analysis is well known from earlier studies (Carvalho and Cohen, 2005; Joty and Hoque, 2016; Mohiuddin et al., 2019). We analyzed our corpus to reveal the sequential relationship. Figure 3a shows the dependencies between adjacent messages and the entropy of a subsequent tag in our corpus. *START* represents a beginning message of a thread. Note that the dependency matrix is a right stochastic matrix and

Speesh acts (BC3/QC3)	Dialogue acts (ours)
Suggestion	*Suggestion*, *Request*
Response	*Commissive*, *Answer*
Question	*Question*
Polite	*Others*
Statement	*Inform*

Table 4: Mapping from our tag set of DAs to speech acts in BC3 and QC3.

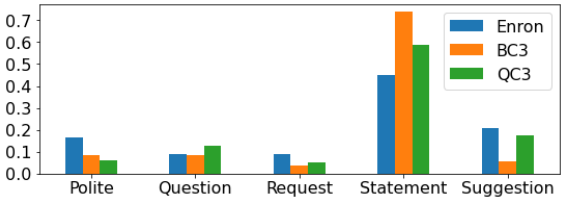


Figure 2: Comparison of tag distributions between our corpus and BC and QC3.

thus each row sums to 1. The lower entropy values of *START* and *Question* indicate a higher correlation between the tag of one message and the tag of a subsequent message. From the dependency matrix of Figure 3a, we can observe that an initial message of a thread is usually a *Question* or *Request* except an *Inform*, and that a *Question* message is most likely to be followed by an *Answer* message. The results agree with our original expectation.

Figure 3b shows the dependencies between a message and a sentence in the message. Note that their probabilities do not sum to 1 in column-wise because a message usually consists of multiple sentences. The higher value of diagonal elements of the matrix indicates that a DA tag of a message corresponds to the same dialogue tag of any one of the sentences in the message. In fact, we found that 90.3% of the messages were annotated with the same dialogue tags as the tags of the sentences in the message. Now, let us consider a scenario in which we can know the ground truth of the sentence-level tags. Under this scenario, we can build a Bernoulli naive Bayes model with sentence-level tags for message-level recognition. The performance of this simple model reaches 70.8% in terms of accuracy. These results indicate a strong correlation between the tag of a message and the tags of sentences in the message.

On the other hand, some of non-diagonal elements of the matrix are moderate. This means that some of sentences in a sentence could have different tags from tag of the message. For example, a *Inform* message contains a *Request* sentence with probability of 0.40. In such cases, annotators (recognition models) need consolidate the sentence-level tags to the message-level tag from the perspective of importance of the communicative functions.

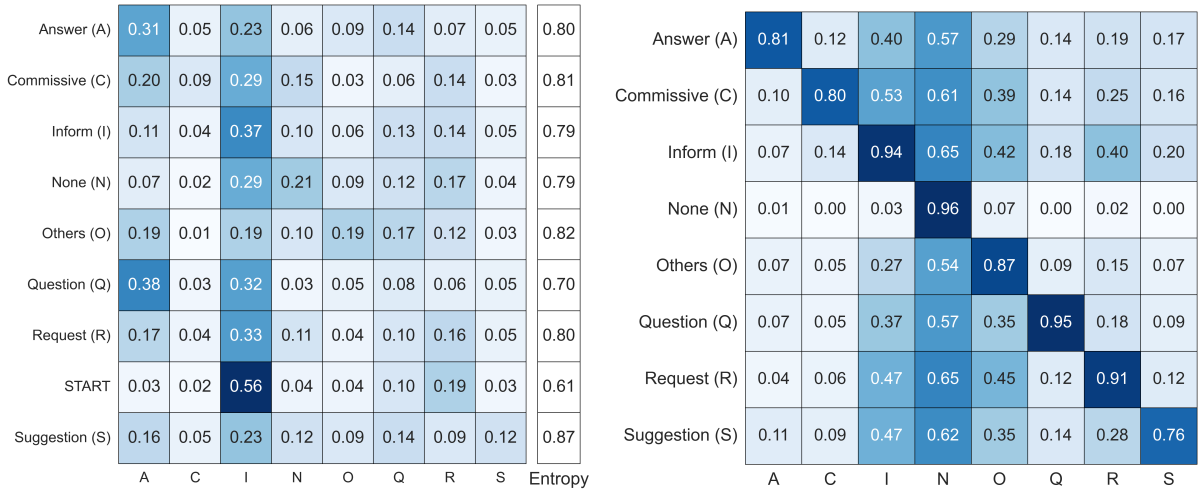


Figure 3: Dependency (Transition) matrices between DA tags of two segments. An element in a matrix represents the probability $P(tag_{column}|tag_{row})$. (a) The dependency matrix between messages. A row represents a DA tag of a message and a column represents a DA tag of the subsequent message. The last column Entropy represents the entropy of the subsequent tag. (b) The dependency matrix between a message and a sentence in the message. A row represents a DA tag of a message and a column represents a DA tag of a sentence in the message.

5 Dialogue Act Recognition

5.1 Models

To assess the difficulty of DA recognition with our corpus, we built several models including handcrafted feature models, existing neural models, and a pre-trained contextual representation model as our baselines. We prepared separate models for sentence- and message-level DA tags.

SVM: We implemented a support vector machine-based model with uni-gram and bi-gram features of a segment. We tuned a soft margin parameter and the number of iterations using the development sets.

LR: We implemented a logistic regression model with uni-gram and bi-gram features of a segment. We tuned a constant of L2 regularization and the number of iterations using the development sets.

CRF: As a classical sequential prediction model, we prepared a linear-chain CRF model with uni-gram and bi-gram features of a segment. A CRF can consider the dependencies of the DA tags in a sequence of segments. We tuned the constants of L1 and L2 regularization, and the number of iterations using the development sets.

LSTM: We prepared bidirectional LSTM model of Joty and Hoque (2016) as a simple neural baseline. The LSTM model encodes a segment to segment representation from the words of the segment.

H-LSTM, H-LSTM-CRF: In Mohiuddin et al. (2019), the authors introduce a neural network model based on a combination of hierarchical LSTMs of two layers and a CRF layer for predicting sentence-level DA tags. In this model, a thread is considered as a sequence of sentences. The first layer of hierarchical LSTMs encodes the sentence representations from word embeddings. The second layer of hierarchical LSTMs updates the sentence representations by considering the surrounding sentences. We also prepared only hierarchical LSTMs of two layers as an H-LSTM. We extended these models to predict message-level DA tags by assuming that a thread is a sequence of messages.

BERT: We fine-tuned the pre-trained contextual representation model BERT (Devlin et al., 2019), which achieved improvements on various natural language processing tasks, using the classification settings. Specifically, we use the uncased BERT-based model[2] as the pre-trained model.

5.2 Experimental Settings

We used the public implementation of Mohiuddin et al. (2019) [3] as the LSTM, H-LSTM, H-LSTM-CRF, and the same hyperparameters as them. In their implementation, the word embedding is pre-trained from the existing corpora of asynchronous conversations by using Glove (Pennington et al., 2014).

We evaluated with a ten-fold cross-validation. We used the annotation data of the crowd workers because of the large amount of data. In each fold, we partitioned the data randomly into a training set (80%), a development set (10%), and a test set (10%). Note that the data is split at the thread level to avoid overlapping threads in different sets. As the evaluation metric, we use a macro-averaged F1 due to the imbalance of the DA tags and report the average values in the ten folds. We choose optimal hyperparamters of models in terms of a macro F1 value in the development set.

5.3 Evaluation Results

Performance at Sentence-Level

The upper half of Table 5 presents sentence-level F1 scores of the baseline models. BERT significantly outperforms the previous best approaches by a margin of 8.5 points. The performance of the handcrafted feature models is comparable. A comparison between the existing neural models shows that there is no difference between LSTM and H-LSTM and that the two models surpass H-LSTM-CRF. These results show that the sequential dependencies of the contexts and tags are not effective in our corpus. The existing neural models have a strong performance compared to the handcrafted feature models, as in previous studies described in Mohiuddin et al. (2019).

Performance at Message-Level

Message-level F1 scores of the baseline models are shown in the lower half of Table 5. Compared with the performance at sentence-level, the performance at message-level is relatively lower. This is because a

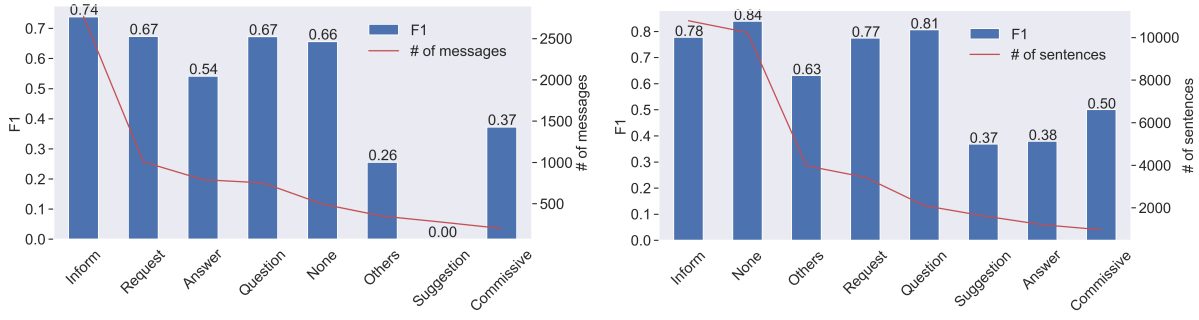
[2]https://github.com/huggingface/transformers

[3]https://ntunlpsg.github.io/project/speech-act/

level	split	SVM	LR	CRF	LSTM	H-LSTM	H-LSTM-CRF	BERT
sentence	dev	0.546	0.545	0.552	0.575	0.575	0.569	**0.647**
	test	0.542	0.538	0.541	0.565	0.565	0.554	**0.650**
message	dev	0.407	0.402	0.400	0.437	0.428	0.428	**0.534**
	test	0.407	0.382	0.366	0.400	0.380	0.368	**0.533**

Table 5: Message-level and sentence-level F1 scores of each models on our corpus. The best scores in each row of the table is highlighted in bold. The best scores in each model category of models (handcrafted feature models and previous neural models) are underlined.

message is more likely to have inherently multiple tags and much harder to find its main communicative function. Similar to the sentence-level, BERT achieves a 13.3 point improvement against the previous best approaches. SVM performs the best among the handcrafted features models. Sequence learning models (CRF, H-LSTM, and H-LSTM-CRF) perform poorly compared to sentence-level predictions. We conjecture that the sequential context information does not contribute effectively because the sequence length of the messages in a thread is shorter than the sequence length of the sentences (2.6 messages vs. 13.3 sentences per thread). The existing neural models slightly underperform the handcrafted feature models unlike the sentence-level predictions. The word length of a message is longer than the word length of a sentence (54.5 vs. 10.5 words), and thus the LSTM layers of the models cannot properly handle the long sequences.



(a) F1 score of BERT at message-level

(b) F1 score of BERT at sentence-level

Figure 4: F1 score of each tags at sentence-level and message-level with the number of tags in our corpus.

6 Discussions

In this section, we further analyze a DA recognition model to provide insight into our corpus. As a target model of the analysis, we choose the BERT model owing to its performance. We use only one of the ten-fold data for the following analysis owing to the computational efficiency.

6.1 Performance of Each Dialogue Act Tags

To further analyze the BERT model, we evaluated the performances of each DA tag. Figure 4 shows the detailed F1 scores of each tag and the frequency of each tag in our corpus. The analysis shows that the performances of less-frequent tags tend to decrease. Specifically, tags of fewer than 500 in number at the message-level are inaccurately predicted.

6.2 Human Performance

We assess the human performance for a comparison with BERT. We used the data of three expert annotators of the held-out set to establish the human performance. More specifically, we treated the data of one

	sentence	message
BERT	0.644 (0.017)	0.544 (0.035)
Experts	0.834 (0.019)	0.728 (0.015)

Table 6: Mean (standard deviation) F1 score of BERT and expert annotators at sentence-levels and message-level on the held-out set.

setup	source	target	BC3	QC3
SELF		✓	0.498	0.387
ZERO-SHOT	✓		0.565	0.599
ZERO-SHOT (REDUCED)	✓		0.226	0.248
TRANSFER	✓	✓	0.686	0.708

Table 7: F1 score on BC3 and QC3. SELF refers to training on the target domain only, ZERO-SHOT refers to training on the source domain only, and TRANSFER refers to training on the source domain and fine-tuning on the target domain.

expert annotator (for instance, Expert 1) as the ground truth and the data of the other expert annotators (Expert 2 and Expert 3) as predictions. This process was followed for each expert annotator. The BERT model was also evaluated on the data of the three expert annotators.

The evaluation results are shown in Table 6. The results show that the expert annotators achieve a significantly higher performance and outperform BERT. This indicates that the DA recognition in our corpus is still a challenge. There are no obvious differences between the BERT performances on the crowd worker data (Table 5) and the expert data (Table 6). Although the annotation quality of the crowd workers is not high, we can regard the performance on the crowd worker data to be a valid evaluation.

6.3 Domain Independence

Our annotation scheme was designed for domain-agnostic tags of the DA. To estimate the domain independence of our annotations, we investigated the performance under three evaluation scenarios. First, a model is trained and evaluated on a target domain (SELF). This scenario is used as a baseline for the other scenarios. Second, a model is trained on a source domain and not trained on the target domain and is evaluated on the target domain (ZERO-SHOT). Finally, a model is trained on the source domain and fine-tuned on the target domain and evaluated on the target domain (TRANSFER).

We used our corpus as a source domain, BC3 or QC3 as the target domain. BC3 and QC3 are conversations of e-mail and online forum, respectively. The DA tags of our corpus are aligned with the five speech tags of BC3 and QC3 using the mapping described in Table 4. Following the previous research of Mohiuddin et al. (2019), the threads of BC3 and QC3 are split randomly into a training set (40%), a development set (20%), and a test set (40%).

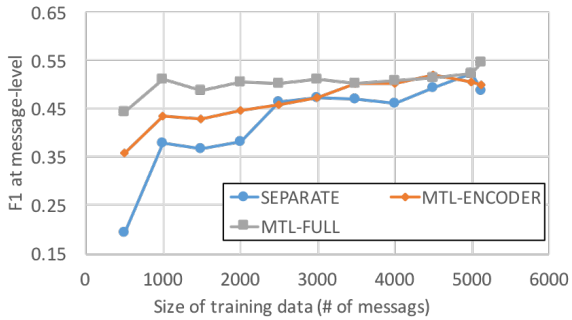
Table 7 shows the F1 score on the three scenarios. Surprisingly, the ZERO-SHOT setup surpasses the SELF setup by 8.7 points in BC3 and 11.2 points in QC3. To analyze the reason for this, we reduced the size of our corpus to the same size as BC3 or QC3 (ZERO-SHOT (REDUCED)). The performance in the ZERO-SHOT (REDUCED) setup decreased significantly and was lower than that of the SELF setup. Therefore, we can suppose that the large size of our corpus clearly contributes to the performance of the target domains. Furthermore, fine-tuning the target domain boosts the performance by 12.1 points for BC3 and 10.9 points for QC3. These results imply that a model trained on our corpus can be generalized for use in other corpora on asynchronous conversations.

6.4 Multi-Task Learning with Sentence-Level and Message-Level Dialogue Acts

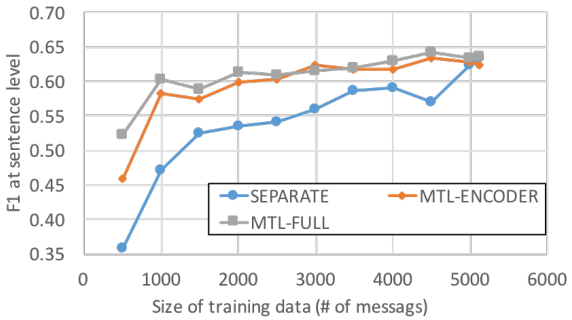
One advantage of our corpus is the two-level annotations of the DA tags. To leverage the annotations for the learning models, we introduce multi-task learning for predicting the sentence- and message-level tags. To estimate the effect of multi-task learning, we conducted experiments on the following three models. First, we train separate BERT models for predicting sentence- and message-level tags, respectively (SEPARATE). These models correspond to the models described in Section 5.3. Second, BERT models are trained by sharing the Transformer encoder layers but the output layers for classification (MTL-ENCODER). Finally, a single BERT model is trained for predicting both sentence- and message-level tags (MTL-FULL). We use a summation of the sentence-level and message-level cross-entropy losses as an objective function of MTL-ENCODER and MTL-FULL.

Figure 5 shows the F1 score for each setting when we increase the number of messages used as

training data. We can see that the performance of all models improves with an increase in the data size. The MTL-ENCODER and MTL-FULL models achieve significant gains over the SEPARATE model when extremely limited training data are available. In particular, the MTL-FULL model achieves a 24.9 point increase in the score F1 at the sentence-level and a 16.6 point increase at the message level as compared with the SEPARATE model when the size of the training data is 500. Comparing the MTL-FULL model with the MTL-ENCODER model, the MTL-FULL model performs the best with a small number of training data. However, the difference between the two models gradually reaches zero as the size of the training data increases. These results show that, although the sentences and messages have different lengths, the MTL-FULL model can accurately classify the sentences and messages in the same manner.



(a) F1 score at message-level DA tags



(b) F1 score at sentence-level DA tags

Figure 5: F1 score of each setup on the development dataset of our corpus. With the SEPARATE setup, the sentence and message classification models are trained separately. With the MTL-ENCODER setup, the encoder of the sentence classification model is shared with the message classification model in training. With the MTL-FULL setup, a single model is trained for both sentence and message classification.

7 Conclusion

Aiming to facilitate the development of DA recognition systems for asynchronous conversations, we developed a large-scale corpus of e-mail conversations with domain-agnostic and two-level DA annotations. A comparison with human and sophisticated neural models demonstrated that there is still plenty of room for advancement in modeling of recognition DA. An evaluation on the domain adaptation shows that training on our corpus contributes to a generalization of the models. We also showed that multi-task learning with two-level tags substantially boosts the performance.

Because an e-mail conversation has a hierarchical structure where words form a sentence, sentences form a message, and messages form a thread, we are planning to explore an efficient end-to-end model that leverages the hierarchical structure. In this paper, we focused annotating communicative functions of DAs. We will extend annotations to other important components of DAs such as rhetorical relations.

Acknowledgments

We would like to thank the anonymous reviewers for their comments to improve and clarify this paper. Computational resource of AI Bridging Cloud Infrastructure (ABCI) provided by National Institute of Advanced Industrial Science and Technology (AIST) was used.

References

Sumit Bhatia, Prakhar Biyani, and Prasenjit Mitra. 2012. Classifying user messages for managing web forum data. In *WebDB*.

Sumit Bhatia, Prakhar Biyani, and Prasenjit Mitra. 2014. Summarizing online forum discussions – can dialog acts of individual messages help? In *Proceedings of the 2014 Conference on Empirical Methods in Natu-*

ral Language Processing (EMNLP), pages 2127–2131, Doha, Qatar, October. Association for Computational Linguistics.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David Traum. 2012. ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 430–437, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Vitor R. Carvalho and William W. Cohen. 2005. On the collective classification of email "speech acts". In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, page 345–352, New York, NY, USA. Association for Computing Machinery.

William W. Cohen, Vitor R. Carvalho, and Tom M. Mitchell. 2004. Learning to classify email into "speech acts". In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 309–316, Barcelona, Spain, July. Association for Computational Linguistics.

Mark G Core and James Allen. 1997. Coding dialogs with the damsl annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, volume 56, pages 28–35. Boston, MA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Rajdip Dhillon, Sonali Bhagat, Hannah Carvey, and Elizabeth Shriberg. 2004. Meeting recorder project: Dialog act labeling guide. Technical report, INTERNATIONAL COMPUTER SCIENCE INST BERKELEY CA.

Donghui Feng, Erin Shaw, Jihie Kim, and Eduard Hovy. 2006. Learning to detect conversation focus of threaded discussions. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 208–215, New York City, USA, June. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

J. J. Godfrey, E. C. Holliman, and J. McDaniel. 1992. Switchboard: telephone speech corpus for research and development. In *Proceedings ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol.1, March.

Liangjie Hong and Brian D. Davison. 2009. A classification-based approach to question answering in discussion boards. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 171–178, New York, NY, USA. Association for Computing Machinery.

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*, pages 1250–1259, Singapore, August. Association for Computational Linguistics.

Shafiq Joty and Enamul Hoque. 2016. Speech act modeling of written asynchronous conversations with task-specific embeddings and conditional structured models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1746–1756, Berlin, Germany, August. Association for Computational Linguistics.

Shafiq Joty and Tasnim Mohiuddin. 2018. Modeling Speech Acts in Asynchronous Conversations: A Neural-CRF Approach. *Computational Linguistics*, 44(4):859–894, December.

Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. 2013. Topic segmentation and labeling in asynchronous conversations. *J. Artif. Int. Res.*, 47(1):521–573, May.

Dan Jurafsky, Liz Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual. Technical report, UC Boulder & SRI International.

Su Nam Kim, Li Wang, and Timothy Baldwin. 2010. Tagging and linking web forum posts. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 192–202, Uppsala, Sweden, July. Association for Computational Linguistics.

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

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Annie Louis and Shay B. Cohen. 2015. Conversation trees: A grammar model for topic structure in forums. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1543–1553, Lisbon, Portugal, September. Association for Computational Linguistics.

Tasnim Mohiuddin, Thanh-Tung Nguyen, and Shafiq Joty. 2019. Adaptation of hierarchical structured models for speech act recognition in asynchronous conversation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1326–1336, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Tatsuro Oya and Giuseppe Carenini. 2014. Extractive summarization and dialogue act modeling on email threads: An integrated probabilistic approach. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 133–140, Philadelphia, PA, U.S.A., June. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Denis Peskov, Nancy Clarke, Jason Krone, Brigi Fodor, Yi Zhang, Adel Youssef, and Mona Diab. 2019. Multi-domain goal-oriented dialogues (MultiDoGO): Strategies toward curating and annotating large scale dialogue data. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4526–4536, Hong Kong, China, November. Association for Computational Linguistics.

Sujith Ravi and Jihie Kim. 2007. Profiling student interactions in threaded discussions with speech act classifiers. In *Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*, page 357–364, NLD. IOS Press.

Emanuel A. Schegloff and Harvey Sacks. 1973. Opening up closings. *Semiotica*, 8(4):289 – 327.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100, Cambridge, Massachusetts, USA, April 30 - May 1. Association for Computational Linguistics.

Jan Ulrich, Gabriel Murray, and Giuseppe Carenini. 2008. A publicly available annotated corpus for supervised email summarization. In *Proc. of AAAI email-2008 workshop, chicago, usa*.

Amy X. Zhang, Bryan Culbertson, and Praveen Paritosh. 2017. Characterizing online discussion using coarse discourse sequences. In *Proceedings of the 11th International AAAI Conference on Weblogs and Social Media*, ICWSM '17.