In Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, Stockholm, Sweden, September 11-13, 2019. pp. 399-410. 2019.

Dialogue Act Classification in Team Communication for Robot Assisted Disaster Response

Tatiana Anakina and Ivana Kruijff-Korbayová
German Research Institute for Artificial Intelligence (DFKI) / Saarbrücken, Germany
tatiana.anikina@dfki.de, ivana.kruijff@dfki.de

Abstract

We present the results we obtained on the classification of dialogue acts in a corpus of human-human team communication in the domain of robot-assisted disaster response. We annotated dialogue acts according to the ISO 24617-2 standard scheme and carried out experiments using the FastText linear classifier as well as several neural architectures, including feed-forward, recurrent and convolutional neural models with different types of embeddings, context and attention mechanism. The best performance was achieved with a "Divide & Merge" architecture presented in the paper, using trainable GloVe embeddings and a structured dialogue history. This model learns from the current utterance and the preceding context separately and then combines the two generated representations. Average accuracy of 10-fold cross-validation is 79.8%, F-score 71.8%.

1 Introduction

Disaster response teams operate in high risk situations and face critical decisions despite partial and uncertain information. First responders increasingly deploy mobile robotic systems to mitigate risk and increase operational capability. In order for robotic systems to provide optimal support for mission execution, they need mission knowledge, i.e., run-time awareness and understanding of the mission goals, team composition, the tasks of the team(s), how and by whom they are being carried out, the state of their execution, etc. Since first responders typically operate under high cognitive load and time pressure, it is paramount to keep the burden of entering mission knowledge into the system at a minimum. The goal of our research thus is to develop methods for *extracting run-time mission knowledge from the verbal communication in the response team*. The acquired mission knowledge can also be used to assist the first responders during or after the mission, for example, by supporting the real-time coordination of human and robot actions or by mission documentation generation (Willms et al., 2019).

In this paper we address one particular subproblem: dialogue act (DA) recognition. DAs are needed for a better understanding of the team communication and how the mission tasks are being executed. For example, *Requests* communicate task assignments and thus allow us to distinguish task assignments from other task-relevant information exchange; *Informs* often report task progress; and *Questions* indicate what was unclear and required additional explanations. These distinctions are also useful for providing assistance, including compiling mission documentation.

We use the corpus of human-human team communication in robot-assisted disaster-response collected in the TRADR project (Kruijff-Korbayová et al., 2015). The TRADR team communication is task-oriented, focused on collaborative execution of a mission by a structured team using mobile robots to remotely gather situation awareness in a complex, dynamic, unknown physical environment. In this the communication differs from that in well-known existing corpora annotated with DAs.

We annotated our corpus with DAs following the ISO 24617-2 scheme (Bunt et al., 2012, 2017) and experimented with several machine learning approaches to DA classification. We explored various models, including different ways of taking dialogue context into account.

We overview previous work on DA classification and existing corpora with DA annotations in §2. We present our corpus in §3 and provide statistics for DA and speaker role distribution. In §4 we describe the classification models tested in our experiments and report the evaluation results. We conclude with a discussion and future plans in §5.

2 Related Work

There is a body of research on teamwork and information sharing in disaster response, with and without robots, e.g., (Casper and Murphy; Burke et al.; Burke and Murphy; Johnson et al., 2017; Toups et al., 2016; Carver and Turoff, 2007).

There has been very little work on dialogue processing in this domain so far. In the pioneering project TRIPS a decision-support dialogue system was developed for the planning of an island evacuation in the event of a natural disaster. Focus was on semantic parsing and task-specific interpretation. This approach was further developed to handle various more complex emergency tasks covered in the Monroe corpus (Stent, 2000). This work focused on mission planning (not execution), data was collected in lab (not real disaster environment) and the participants were students (not real first responders). DAs were annotated using the DAMSL scheme (Core and Allen, 1997).

Some works on human-robot collaboration for disaster response address the interpretation of verbal commands to robots (Kruijff et al., 2014; Yazdani et al., 2018), but not the overall team communication.

In (Martin and Foltz, 2004) automatic analysis of the semantic content of team communication and automatic verbal behavior labeling was used to assess team performance in a command and control task with an unmanned aerial vehicle in a simulated environment. A corresponding synthetic team-member agent is described in (Cooke et al., 2016). Since the corpus is not available and the publications do not provide details on the task and communication complexity, a closer comparison to our work is not possible. Communication analysis was used also in (Burke et al.). They designed and manually applied a team communication coding scheme, in order to examine robot operator situation awareness and technical search team interaction during a high-fidelity disaster response drill with teleoperated robots. DAs are reflected in their annotation of the forms and functions of communication contributions.

Corpora with DA annotations include also well-known human-human dialogue corpora, such as MapTask (Anderson et al., 1991; Carletta et al., 1997); TRAINS (Allen, 1991); Switchboard (Godfrey et al., 1992); Meeting Recorder Dialogue Act (Shriberg et al., 2004) and the AMI Meeting Corpus (Carletta et al., 2005), and recent large corpora, e.g., Maluuba Frames (Schulz et al., 2017) and MultiWOZ (Budzianowski et al., 2018)). These corpora cover different domains and the goals the participants follow in their interaction are quite different from what is going on in the team communication in our domain.

Despite the differences it would be interesting to see how DA classification models developed on other exiting corpora perform on our corpus. The challenge of such endeavor is, however, that different and sometimes very task-specific schemes have been applied to annotate DAs. For instance, some of the DAs in the Maluuba Frames corpus include domain-specific labels such as *Canthelp* and *No_result* as well as *Thankyou* and *Moreinfo*.

The ISO 24617-2 standard for DA annotations introduced in (Bunt et al., 2012) and further defined in (Bunt et al., 2017) was proposed to overcome this. To date several corpora have been annotated accordingly and made available through the DialogBank (Bunt et al., 2016). Although the mapping of DA labels from other annotations to the ISO standard is quite straightforward in some cases (e.g., for *Inform* or *Request*), in other cases the specificity of the domain prevents from further generalizations, as discussed in (Chowdhury et al., 2016). These issues lead us to postpone transfer learning for future work and start traditionally by experiments on our own corpus.

Previous work on automatic DA classification includes the use of Hidden Markov models (Stolcke et al., 2000), Maximum Entropy (Choi et al., 1999), Generative and Conditional Bayesian Networks (Ji and Bilmes, 2005), and Support Vector Machines (Quarteroni and Riccardi, 2010). Recent papers also explored neural architectures (Kumar et al., 2017; Liu et al., 2017) and compared word embeddings (Cerisara et al., 2018).

Only few works to date systematically tested different kinds of context for DA classification. Several experiments on the Switchboard corpus are described in (Ribeiro et al., 2015), which tested untagged and index-tagged n-grams as well as context presented in the form of dialog act annotations for the previous segments. Index-tagged n-grams (n-grams tagged with the distance to the current segment) improved accuracy significantly, from 70.6% to 75.1%, and the DA annotations for the preceding segments even to 76.4%.

(Liu et al., 2017) tested different kinds of context for DA classification using deep neural mod-

els. They present hierarchical models based on convolutional neural networks (CNN) for sentence representations which they combine with dialogue history. They encode context as previous DA labels and as probabilities for system predictions, and experiment with dialogue history of varied length. Including context information in their models evaluated on the Switchboard corpus resulted in significant increase of accuracy from 77% to almost 80%. These results indicate that context should be taken into account when processing structured conversations.

3 The Corpus

We use the corpus of robot-assisted disaster-response team communication collected during joint exercises with first responders in the TRADR project (Kruijff-Korbayová et al., 2015).[1] The TRADR corpus contains audio recordings and transcriptions of the speech communication in a team of firefighters using robots in the aftermath of an incident, e.g., an explosion, at an industrial site. The team members have various roles: mission commander (MC), team leader (TL), operators (OP) of multiple ground (UGV) and aerial (UAV) robots. They explore the site, searching for persons, hazard sources, fires and other relevant points of interest. The MC and the TL lead the mission. They request situation information from the OPs, who report back with updates and can also share photos taken by the robot camera (see the example in Appendix A).

The recordings were collected during several field tests in 2015, 2016 and 2017. They amount to approximately 10 hours and contain almost 3k speech turns (see Table 1 for details). The 2015 and 2016 recordings are in German, the 2017 ones in English. For the experiments presented in this paper we used the original English data as well as English translations from German. We started on English because of available resources.

Before annotating DAs following the ISO 24617-2 scheme (Bunt et al., 2012, 2017). we segmented the data into *utterances*; we split and merged some turns to obtain appropriate spans for assigning DAs. This resulted in 2469 utterances.

The ISO scheme defines several dimensions and for each of them a hierarchy of communicative functions (a.k.a. DAs). The first author and another annotator independently annotated each utterance with one of the dimensions and a corresponding DA. Inter-annotator agreement was κ=.77 for dimension assignment and κ=.55 (weighted κ=.66) for the generic communicative functions in the *Task* dimension. For the experiments in this paper we used the first author's annotations as a golden reference. We focused on the classification of DAs from the dimensions *Task*, *Feedback* and *Turn Management* (see Table 2 for the used labels).

We annotated the corpus in full compliance with the ISO scheme. Since some DAs had too few occurrences in the corpus we used a simplified set of DA labels in the experiments (see §4.1). The simplified labels are a result of a direct mapping from the ISO scheme labels (see Table 2), making it easy to compare DA classification results. In most cases the simplified labels can be seen as

Recording	Mission	Duration	Turns
TJex 2015			**374**
	Day 1	48:21 min	201
	Day 2	33:21 min	173
TEval 2015			**1165**
	Day 1	58:23 min	289
	Day 2	65:04 min	299
	Day 3	57:15 min	219
	Day 4	53:22 min	358
TEval 2016			**421**
	Day 1	n.a.	311
	Day 2	n.a.	110
TEval 2017			**822**
	Day 1	64:02 min	240
	Day 2	149:20 min	408
	Day 3	56:36 min	174
Total:			**2782**

Table 1: Corpus composition

ISO Annotation Label	Classification Label
Turn Management	Contact
Inform, Promise, Offer, Address-Suggestion	Inform
PositiveFeedback, AcceptRequest, AcceptOffer, AcceptSuggestion, Agreement	Affirmative
Request	Request
CheckQuestion, SetQuestion, ChoiceQuestion, Question	Question
Confirm	Confirm
Disconfirm	Disconfirm
Negative Feedback, DeclineOffer, Disagreement	Negative

Table 2: Mapping of ISO annotation labels to labels for automatic classification

[1]The TRADR team communication corpus is available online from www.tradr-project.eu/resources/datasets/ or talkingrobots.dfki.de/resources/tradr/

Dialogue Act	MC	TL	OP	Total
Contact	**32**	**350**	360	742
Inform	19	132	**476**	627
Affirmative	8	217	127	352
Request	9	262	3	274
Question	12	150	84	246
Confirm	2	28	131	161
Disconfirm	0	4	49	53
Negative	0	6	8	14

Table 3: Dialogue act distribution

generalized ISO DAs which were selected based on their utility for the disaster response domain.

The mission interactions consist of *threads*, which are dialogue sequences where two (occasionally multiple) team members talk about a task or situation update, e.g., the TL talks to an OP as illustrated in the example in Appendix A. A new thread is initiated by establishing contact following the standard radio communication protocol. The threads are a good candidate for dialogue context and we used thread history in some experiments as we will describe in the next sections.

4 Experiments

4.1 Pre-processing

Before running the experiments we pre-processed the data as follows.

First, we collapsed DA labels which had very low frequency in the corpus with more frequent ones. For instance, there were only 2 cases of *AddressSuggestion* and 9 cases of *AcceptOffer* in total. Low frequency labels would introduce noise and prevent the classifier from learning reliable patterns. Moreover, there were some ambiguous cases with several possible annotations (e.g. *Inform* and *Promise* for *"I'll send it over to you"*) and we decided to retain the most frequent label to reduce the perplexity. Table 2 shows the mapping of the manually annotated ISO scheme labels to the DA labels used for the automatic classification. The resulting distribution of DA labels is shown in Table 3.

Second, we removed all punctuation. Although punctuation can be a good clue for some DAs (e.g., *"?"* usually indicates *Question*) we removed it, because the ASR software often does not provide punctuation reliably. We also transformed all texts to lower case and padded sequences when using neural networks. For 10-fold cross-validation we split the 2469 utterances into 2222 for training and 247 for testing in each fold partition.

4.2 Baselines

We implemented three baselines. The **majority baseline** assigned each utterance the most frequent label for the given role, i.e., all MC/TL utterances were annotated as *Contact* and all OP utterances as *Inform*. This resulted in accuracy 34.8%

The fact that all TL utterances were classified as *Contact* was an obvious drawback. We therefore tried a **relative-frequency baseline** as an alternative, using the relative frequencies for each DA on the complete corpus (cf. Table 3). Each utterance was assigned a random class based on the relative frequencies. This baseline had accuracy 24.7%[2]. The majority baseline which used solely the role was substantially better compared to the frequency-based random baseline.

The third **mixed baseline** was based on the assumption that all instances of *Contact* are identified correctly and for all other utterances we used the majority baseline. Therefore, the third baseline assigned *Request* to all MC/TL utterances and *Inform* to all OP utterances which were not labeled as *Contact*. This baseline had accuracy 47.2%. Since these three baselines had such a low performance we considered the results of the FastText classifier as a baseline for evaluating the performance of the neural models.

4.3 FastText

As the first model for DA classification we tested FastText[3], an open-source library for text classification and representation using supervised learning with multinomial logistic regression. Although it can represent input text in the form of embeddings it belongs to the family of linear classifiers. We ran FastText using the parameters recommended for a small training set (10 dimensions, 0.5 learning rate, 20 epochs). The average accuracy over a 10-fold cross-validation was 74.0%. It was consistent across the folds (see Table 4). Because of the strong correlation between the speaker role and the DA distribution, as shown in Table 3, we also experimented with including the role as a special token at the beginning of each utterance. This additional information improved the average accuracy to 75.6% and also the accuracy in most folds (see Table 4). Finally, we tested the effect of adding the dialogue thread con-

[2]We also tested a baseline based on DA relative frequencies per role, but the accuracy was even lower, 21%.

[3]https://fasttext.cc/

Fold	Accuracy without Role	Accuracy with Role	Accuracy with Role + Thread History
1	0.656	**0.668**	0.628
2	0.607	**0.684**	0.583
3	0.668	**0.696**	0.583
4	0.745	**0.757**	0.583
5	0.834	**0.858**	0.692
6	**0.761**	0.741	0.640
7	**0.794**	0.773	0.709
8	0.781	**0.785**	0.660
9	0.769	**0.794**	0.676
10	0.785	**0.801**	0.650
Avg:	0.740	**0.756**	0.640

Table 4: FastText 10-fold cross-validation

text: we appended the corresponding thread history to each utterance and trained FastText on this extended input. Accuracy dropped for all folds, to 64.0% on average as shown in Table 4.

4.4 Neural Networks

Neural networks have already shown great potential in tackling various NLP tasks, including DA classification (Chen et al., 2018; Liu et al., 2017). We therefore also tested various neural architectures to classify DAs in our corpus: Feed-Forward Neural Networks (FFNN); Recurrent Neural Networks (RNN), in particular Long-Short Term Memory (LSTM) and bidirectional LSTM models; Convolutional Neural Networks (CNN). We experimented with attention and different kinds of embeddings (including Word2Vec, GloVe and FastText). We also tested the effect of the dialogue context in the form of the preceding thread history concatenated with the current utterance. We present the models and the DA classification results in the next sections.

Feed-Forward Neural Networks

We implemented a simple FFNN using the Keras[4] library with one Embedding layer (we experimented with 100, 200 and 300 dimensions) and applied global average pooling to average the embeddings of all words in the utterance before sending them through the Dense layer. The architecture is shown in Figure 1.

We set the minibatch size to 8, trained the network for 5 epochs and used Adam as an optimizer. We trained several models using the Embedding layer provided by Keras as well as pre-trained GloVe embeddings obtained from the Stanford

[4] https://keras.io/

Embeddings Type	Accuracy
Keras 100	0.755
Keras 200	0.761
Keras 300	**0.762**
GloVe 100, frozen	0.685
GloVe 200, frozen	0.711
GloVe 300, frozen	**0.722**
GloVe 100, trainable	0.759
GloVe 200, trainable	0.768
GloVe 300, trainable	**0.771**

Table 5: DA classification results for FFNNs with different types of embeddings

NLP group website,[5] which were learnt on the data from Wikipedia 2014 and Gigaword 5 (6B tokens, 400K vocabulary). We also experimented with both frozen and trainable embeddings. The results were consistently better with trainable embeddings compared to the frozen version. Table 5 shows the evaluation results with accuracy scores averaged across 10 folds.

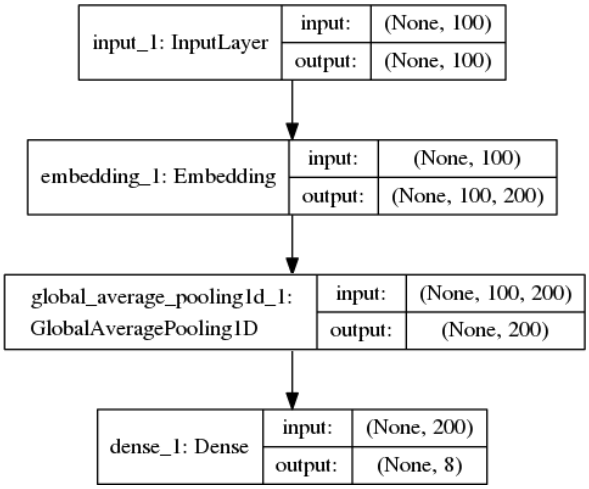

Figure 1: Feed-Forward Network with embeddings [6]

Convolutional Neural Networks

Inspired by the results on DA classification with CNNs in (Liu et al., 2017) we also tested CNNs with varying number of convolutional layers and filter sizes on our data. Figure 2 shows a sample architecture with two convolutions and 128 filters of size 5. We also tested CNNs with different embeddings. The best performance (average accuracy 72.1%) was achieved by the model with one convolutional layer, filter size 10 and embeddings

[5] https://nlp.stanford.edu/projects/glove/

[6] None is a dynamic length dimension which means that a corresponding layer can have variable-length sequences as an input.

Embeddings Type	Conv.	Filter Size	Accuracy
Keras 100	2	5	0.685
Keras 200	2	5	0.697
Keras 100	1	10	**0.721**
Keras 200	1	10	0.712
GloVe 100	2	5	0.695
GloVe 200	2	5	0.694
GloVe 100	1	10	0.703

Table 6: DA classification results for CNN models

trained on our data with dimensionality 100. An overview of the results obtained with various CNN architectures is in Table 6. Interestingly, more complex models resulted in worse scores. Convolutions appear not very useful for the relatively short texts of dialogue utterances.

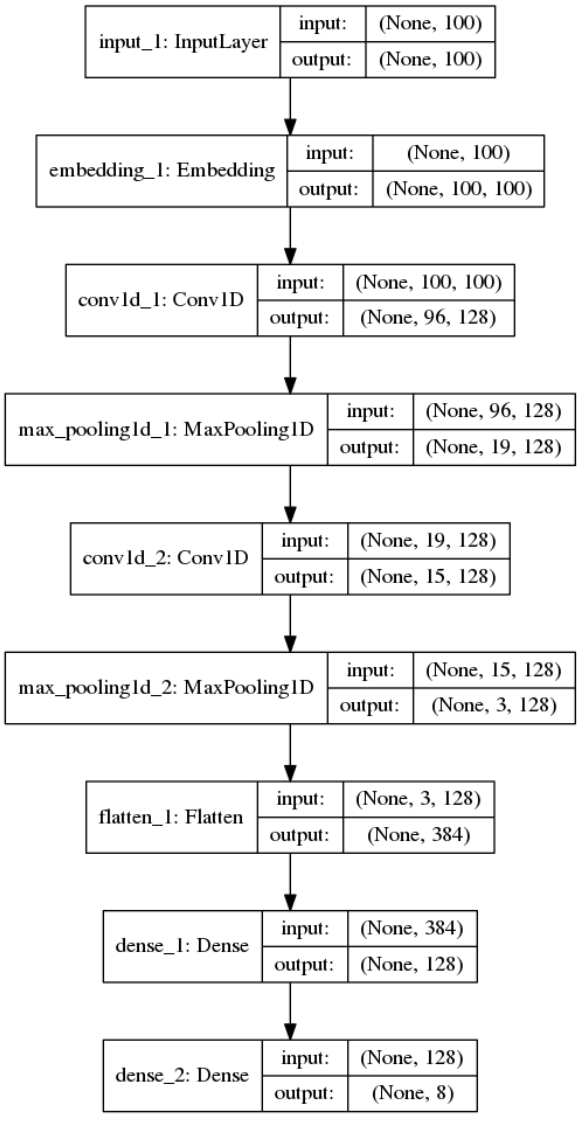

Figure 2: Convolutional Neural Network [6]

Model	Embeddings Type	Accuracy
LSTM	Embedding 200	0.745
LSTM	GloVe 200	0.775
LSTM	GloVe 300	0.768
LSTM +Attention, -Thread	Embedding 200	0.676
LSTM +Attention, -Thread	GloVe 200	0.767
LSTM -Attention, +Thread	GloVe 200	**0.780**
LSTM +Attention, +Thread	GloVe 200	0.745

Table 7: RNN performance

Model	Embeddings Type	Accuracy
no LSTM	GloVe 200	0.784
LSTM for turn & thread	GloVe 200	0.768
LSTM for turn	GloVe 200	**0.798**
LSTM for turn	Word2Vec 100	0.769
LSTM for turn	Word2Vec 200	0.773
LSTM for turn	Word2Vec 300	0.767
LSTM for turn	FastText 300	0.770

Table 8: Divide&Merge performance

Recurrent Neural Networks

We tested RNNs with Long Short Term Memory (LSTM) cells, both LSTMs and bidirectional LSTMs. We also applied an attention mechanism and experimented with various embeddings and regularization parameters. In some experiments we concatenated all previous utterances from the same thread with the current utterance in order to give more context to the classifier. We inserted a *#START#* symbol between the current utterance and the thread text as a separator.

Figure 3 shows the RNN architecture with bidirectional LSTM and attention mechanism. The attention layer follows the idea proposed in (Raffel and Ellis, 2015). We passed the generated word vectors through bidirectional LSTM and multiplied the input with the attention vector at each time step. The result was passed through the Dense layer with ReLU as an activation function. Dropout 0.25 was applied to the function output before it went through the final Dense layer. We tested this model with single utterances as well as with utterances concatenated with their corresponding thread history, with and without attention. The results of different RNN architectures are in Table 7. The best accuracy of 78.0% was achieved by the model which used the thread history and pre-trained GloVe embeddings with trainable weights, no attention.

In the experiments described above we noticed

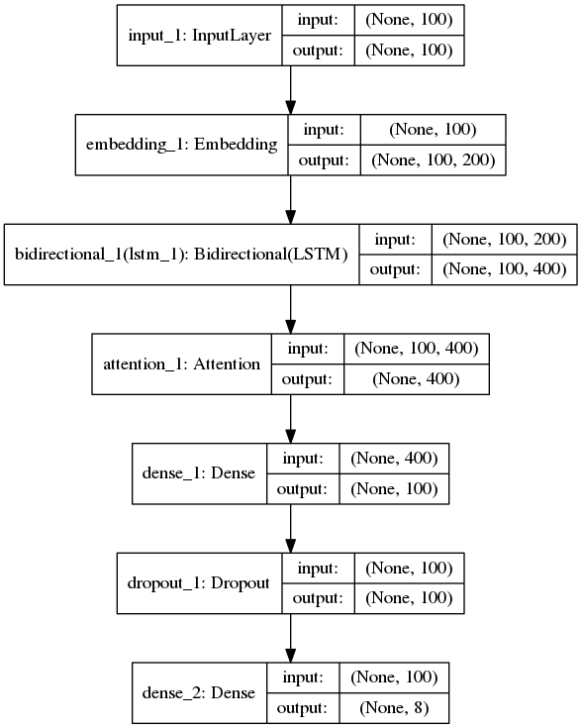

Figure 3: RNN with attention [6]

that simple concatenation of the current utterance with the previous context gives us a very small improvement in accuracy compared to the model which does not use the thread history (accuracy increased from 77.5% to 78.0%). The network treats the current utterance and the thread history as a single input, and this might result in a sub-optimal representation. Hence, we designed a model that learns from the current utterance and from the previous context separately and then combines the two generated representations into one. Because it first separates the current utterance from the context and then puts the representations together we call this new model *Divide & Merge* (D&M). Figure 4 shows the D&M model architecture we implemented. 10-fold cross-validation yielded the best average accuracy of 79.8% using pre-trained GloVe embeddings with 200 dimensions and training for 5 epochs. Detailed results of the D&M model evaluation are in Tables 8 and 9.

Table 8 shows the results for various experimental settings. First, we report the accuracy scores obtained by the D&M model without LSTM, D&M which uses LSTM for encoding both turn and thread utterances and D&M which uses LSTM only for turns while the thread information is encoded using one Embedding layer and global average pooling as shown in Figure 4. The model with turn-only LSTM achieved the best accuracy 79.8%. Second, we also compared different word embeddings (GloVe, Word2Vec and FastText) and found that pre-trained GloVe embeddings with 200 dimensions work best on our data.

Fold	Accuracy
1	0.733
2	0.717
3	0.765
4	0.794
5	0.834
6	0.826
7	0.858
8	0.810
9	0.818
10	0.829
Avg:	**0.798**

Table 9: Divide&Merge 10-fold cross-validation

4.5 Discussion

To compare the performance of the D&M model (accuracy 79.8%) against that of the FastText classifier (accuracy 75.6%) we applied a randomized test with 10,000 trials. The resulting p-value of 0.0001 indicates a significant difference. The accuracy of both FastText and D&M is also significantly better than that of the baselines (24.7% for the relative-frequency baseline, 34.8% for the majority baseline and 47.2% for the mixed baseline). Table 10 contains the results for precision, recall and F-score per DA.

	FastText			Divide&Merge		
Category	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
Contact	0.94	0.96	0.95	0.96	0.98	0.97
Inform	0.70	0.77	0.74	0.75	0.78	0.76
Affirmative	0.78	0.80	0.79	0.81	0.82	0.82
Request	0.69	0.68	0.68	0.75	0.76	0.75
Question	0.58	0.54	0.56	0.71	0.61	0.65
Confirm	0.40	0.28	0.33	0.48	0.50	0.49
Disconfirm	0.60	0.51	0.55	0.60	0.55	0.57
Average (w/o Neg.):	**0.67**	**0.65**	**0.66**	**0.72**	**0.71**	**0.72**
Average (with Neg.):	**0.59**	**0.57**	**0.57**	**0.63**	**0.62**	**0.63**

Table 10: FastText and D&M results per DA

We also compared the performance of the D&M model with threads to the same model without thread information. The results are in Table 11. Note that Tables 10 and 11 show average precision, recall and F1 score for two cases: with and without the category *Negative*. *Negative* turned out to be very difficult to classify because of the

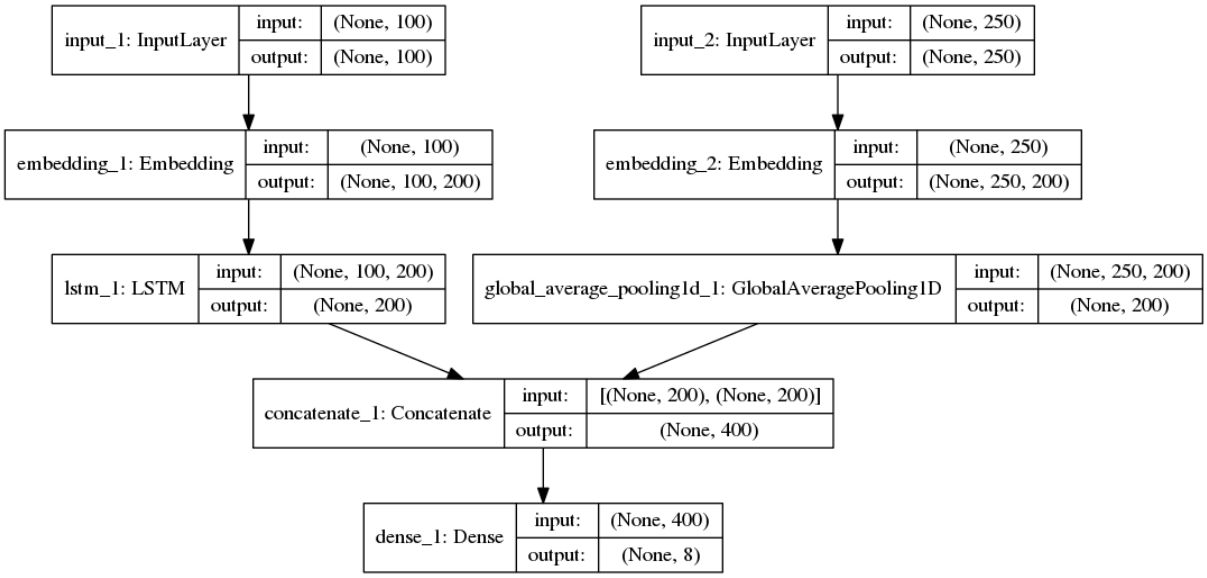

Figure 4: Divide&Merge architecture [6]

following reasons. First, there is a data sparsity problem, because *Negative* has only 14 occurrences in the whole corpus. Second, *Negative* is very similar to *Disconfirm* and in many cases they can be used interchangeably. However, *Negative* was omitted only in the precision and recall calculations showing performance per DA. All accuracy scores presented in this paper take *Negative* into consideration.

Table 11 shows that F1 score increases when the thread information is provided as an additional input to the model. For all DAs except for *Disconfirm* and *Negative* we observe an improvement in terms of precision, recall and F1 score. The poor performance of D&M model on categories *Negative* and *Disconfirm* could be due to the fact that some threads are interconnected and *Negative* is often a response to the previous thread.For instance, in one thread the OP says *"I will put snapshots in ..."* And in the next thread the TL says *"I don't have snapshots"* which should be interpreted as *Negative* with respect to the previous statement. However, D&M classifies the utterance as *Inform* because it does not see the connection between two different threads.

Further manual checking of the classification results confirmed that the D&M model could handle DAs which depend on the context better. Table 12 illustrates this: In Thread 1 FastText almost always picked *Inform* as the most likely label, whereas D&M assigned more DAs correctly. In Thread 2 FastText assigned *Contact* for *"Yeah, I am driving closer now"*. Although there were some instances of *Contact* in the training corpus starting with *"yeah"*, *Contact* is not a good candi-

	D&M no threads			**D&M with threads**		
Category	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
Contact	0.95	0.96	0.95	0.96	0.98	0.97
Inform	0.74	0.73	0.73	0.75	0.78	0.76
Affirm.	0.80	0.76	0.78	0.81	0.82	0.82
Request	0.73	0.74	0.74	0.75	0.76	0.75
Question	0.64	0.60	0.62	0.71	0.61	0.65
Confirm	0.37	0.47	0.41	0.48	0.50	0.49
Disconfirm	0.62	0.59	0.60	0.60	0.55	0.57
Negative	0.25	0.07	0.11	0.00	0.00	0.00
Average (w/o Neg.):	**0.69**	**0.69**	**0.69**	**0.72**	**0.71**	**0.72**
Average (with Neg.):	**0.64**	**0.61**	**0.62**	**0.63**	**0.62**	**0.63**

Table 11: D&M results with and without threads

Speak.	**Text**	**FastText**	**D&M**
Thread 1			
TL	*UGV 1 to team leader.*	**Contact**	**Contact**
OP	*I am coming.*	Inform	**Contact**
TL	*Can you find out whats standing in all this smoke?*	Inform	**Question**
OP	*Yes. I could. You should have a picture of that.*	Inform	**Confirm**
TL	*I'll check that.*	**Inform**	Affirm.
Thread 2			
TL	*Can you get closer to the blue barrel, so that we can see the label?*	**Request**	**Request**
OP	*Yeah, I am driving closer now.*	Contact	**Affirm**

Table 12: Sample DA classification results by FastText and D&M. Correctly assigned DAs are typeset in bold.

date in this case given that the previous utterance was labeled as *Request*. This shows that thread history has an impact on the output of the D&M model. The D&M model makes better use of the thread history than FastText and seems to offer a better model for structured conversations.

In general, the independence assumption made by FastText impairs the classification performance. However, adding thread history resulted in an accuracy drop from 75.6% to 64.0% (cf. §4.3). This means that it is not only thread information that is important for correct classification but also the way this information is encoded and processed by the classifier. Whereas FastText treats the current utterance and the thread history in a bag-of-words fashion, the D&M model treats them as two independent inputs which are being processed by two different parts of the network and their representations are concatenated only at the final stage.

We also tested several models on the part of the Switchboard Corpus available in DialogBank (Bunt et al., 2016). After pre-processing similar to what we did for our corpus we had 443 utterances. We split them into 333 (75%) for training and 110 (25%) for testing. FastText achieved accuracy 60%. Among the neural models a simple FFNN using the Embedding layer initialized with pre-trained GloVe embeddings with 100 dimensions achieved best accuracy 73.6%. The D&M model could not be applied to the DialogBank-Switchboard data because there are no clearly delimited threads. It would be interesting to test the D&M approach on other corpora with dialogues structured into threads similarly to our corpus.

5 Conclusions

We presented the results of dialogue act classification in robot-assisted disaster response team communication. We experimented with a FastText classifier and various neural models using FFNNs, RNNs and CNNs with different types of embeddings and context information, with and without attention. We found that including the speaker role is beneficial whereas adding the previous sentence as dialogue context leads to lower accuracy. This might be due to the fact that dialogues in our corpus consist of threads and concatenating an utterance with a preceding one from a different thread causes erroneous predictions. We then designed the Divide&Merge model, where we added thread history in a separate layer and concatenated not texts but their vector representations. This resulted in a significant improvement with average accuracy 79.8%. Using LSTM cells was beneficial for utterance encodings but the thread history was better encoded using the Embedding layer and global average pooling. Pre-trained GloVe embeddings with dimensionality 200 performed best on our data and the results were slightly better with trainable embeddings. This could be due to the fact that in our corpus some words have non-standard interpretations based on the communication protocol (e.g., *"roger that"*), which are learned from the corpus when we use trainable embeddings.

Incorporating thread information significantly improved DA classification. In the future we wish to investigate more the nature and importance of threads in team communication, e.g., whether to model threads implicitly (as we did) or explicitly; how to best segment them; how important is it to represent intertwined threads; is information throughout a thread used for interpretation or is the influence more local at the thread boundary.

In future work we will also apply the models presented here on the German data in the TRADR corpus; test their performance on the outputs of ASR without any editing by human annotators; look for ways to further improve performance, e.g., by enlarging the corpus by adding relevant dialogues from other corpora. We will develop models for the recognition of mission tasks and distinguishing task requests and commitments by the team members from other task mentions. We will then combine dialogue act and task recognition in a single model. We will release the corpus with the ISO dialogue act annotations later this year.

The models we develop are being integrated as part of the speech processing pipeline in a mission-support system that provides process assistance and facilitates the creation of mission documentation (Willms et al., 2019). It will be evaluated in practice with and by first responders.

Acknowledgements

This work is part of the A-DRZ project funded by the German Ministry of Education and Research (BMBF), grant No. I3N14856.[7] We wish to thank Stefania Racioppa and Natalia Skachkova for their contributions to annotate the TRADR corpus and all our colleagues for valuable discussions.

[7] A-DRZ (Setup of the German Rescue Robotics Center). URL: rettungsrobotik.de

References

James F. Allen. 1991. Discourse structure in the TRAINS project. In *Speech and Natural Language, Proceedings of a Workshop held at Pacific Grove, California, USA, February 19-22. 1991*.

Anne Anderson, M. Bader, Ellen Bard, E. Boyle, Gwyneth M. Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, J. McAllister, J. Miller, Cathy Sotillo, Henry Thompson, and R. Weinert. 1991. The HCRC Map Task corpus. *Language and Speech*, 34(4):351–366.

Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 5016–5026.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Kôiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David R. 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 2012, Istanbul, Turkey, May 23-25, 2012*, pages 430–437.

Harry Bunt, Volha Petukhova, and Alex Chengyu Fang. 2017. Revisiting the ISO standard for dialogue act annotation. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)*.

Harry Bunt, Volha Petukhova, Andrei Malchanau, Kars Wijnhoven, and Alex Chengyu Fang. 2016. The dialogbank. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*.

Jennifer Burke, Michael D. Coovert, Robin Murphy, and Dawn L. Riddle. Moonlight in Miami: An ethnographic study of human-robot interaction in USAR. *Human-Computer Interaction, special issue on Human-Robot Interaction*, 19:85.

Jennifer L. Burke and Robin Murphy. From remote tool to shared roles. *IEEE Robotics and Automation Magazine, special issue on New Vistas and Challenges for Teleoperation*, 15:39.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2005. The AMI meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction, Second International Workshop, MLMI 2005, Edinburgh, UK, July 11-13, 2005, Revised Selected Papers*, pages 28–39.

Jean Carletta, Stephen Isard, Gwyneth Doherty-Sneddon, Amy Isard, Jacqueline C. Kowtko, and Anne H. Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31.

Liz Carver and Murray Turoff. 2007. Human-computer interaction: The human and computer as a team in emergency management information systems. *Communications of the ACM*, 50(3):33–38.

Jennifer Casper and Robin Murphy. Human-robot interaction during the robot-assisted urban search and rescue response at the World Trade Center. *IEEE Transactions on Systems, Man and Cybernetics Part B*, 33:367.

Christophe Cerisara, Pavel Král, and Ladislav Lenc. 2018. On the effects of using word2vec representations in neural networks for dialogue act recognition. *Computer Speech & Language*, 47:175–193.

Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018. Dialogue act recognition via crf-attentive structured network. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 225–234.

Won Seug Choi, Jeong-Mi Cho, and Jungyun Seo. 1999. Analysis system of speech acts and discourse structures using maximum entropy model. In *27th Annual Meeting of the Association for Computational Linguistics, University of Maryland, College Park, Maryland, USA, 20-26 June 1999*.

Shammur Absar Chowdhury, Evgeny A. Stepanov, and Giuseppe Riccardi. 2016. Transfer of corpus-specific dialogue act annotation to ISO standard: Is it worth it? In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*.

Nancy J Cooke, Mustafa Demir, and Nathan McNeese. 2016. Synthetic teammates as team players: Coordination of human and synthetic teammates. Technical report, Cognitive Engineering Research Institute, Mesa (US).

Mark G. Core and James F. Allen. 1997. Coding dialogues with the DAMSL annotation scheme. In *Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35, Menlo Park, California. AAAI, American Association for Artificial Intelligence.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP'92, pages 517–520, Washington, DC, USA. IEEE Computer Society.

Gang Ji and Jeff A. Bilmes. 2005. Dialog act tagging using graphical models. In *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '05, Philadelphia, Pennsylvania, USA, March 18-23, 2005*, pages 33–36.

Matthew Johnson, Brandon Shrewsbury, Sylvain Bertrand, Duncan Calvert, Tingfan Wu, Daniel Duran, Douglas Stephen, Nathan Mertins, John Carff, William Rifenburgh, and Jesper Smith. 2017. Team ihmcs lessons learned from the darpa robotics challenge: Finding data in the rubble. *Journal of Field Robotics*, 34(2):241.

Geert-Jan M. Kruijff, Ivana Kruijff-Korbayová, Shanker Keshavdas, Benoit Larochelle, Miroslav Janíček, Francois Colas, M. Liu, Francois Pomerleau, Roland Siegwart, Mark A. Neerincx, Rosemarijn Looije, Nanja J.J.M Smets, Tina Mioch, Juriaan van Diggelen, Fiora Pirri, Mario Gianni, F. Ferri, M. Menna, Rainer Worst, T. Linder, V. Tretyakov, Hartmut Surmann, Tomáš Svoboda, Michael Reinštein, Karel Zimmermann, Tomáš Petříček, and Václav Hlaváč. 2014. Designing, developing, and deploying systems to support humanrobot teams in disaster response. *Advanced Robotics*, 28(23):1547–1570.

Ivana Kruijff-Korbayová, Francis Colas, Mario Gianni, Fiora Pirri, Joachim de Greeff, Koen Hindriks, Mark Neerincx, Petter Ögren, Tomáš Svoboda, and Rainer Worst. 2015. Tradr project: Long-term human-robot teaming for robot assisted disaster response. *KI - Künstliche Intelligenz*, 29(2):193–201.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, Sachindra Joshi, and Arun Kumar. 2017. Dialogue act sequence labeling using hierarchical encoder with CRF. *CoRR*, abs/1709.04250.

Yang Liu, Kun Han, Zhao Tan, and Yun Lei. 2017. Using context information for dialog act classification in DNN framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2170–2178.

Melanie J Martin and Peter W Foltz. 2004. Automated team discourse annotation and performance prediction using LSA. In *Proc. of HLT-NAACL 2004: Short Papers*, pages 97–100. ACL.

Silvia Quarteroni and Giuseppe Riccardi. 2010. Classifying dialog acts in human-human and human-machine spoken conversations. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 2514–2517.

Colin Raffel and Daniel P. W. Ellis. 2015. Feed-forward networks with attention can solve some long-term memory problems. *CoRR*, abs/1512.08756.

Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. 2015. The influence of context on dialogue act recognition. *CoRR*, abs/1506.00839.

Hannes Schulz, Jeremie Zumer, Layla El Asri, and Shikhar Sharma. 2017. A frame tracking model for memory-enhanced dialogue systems. In *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Vancouver, Canada, August 3, 2017*, pages 219–227.

Elizabeth Shriberg, Rajdip Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the SIGDIAL 2004 Workshop, The 5th Annual Meeting of the Special Interest Group on Discourse and Dialogue, April 30 - May 1, 2004, Cambridge, Massachusetts, USA*, pages 97–100.

Amanda Stent. 2000. The monroe corpus. Technical report, Dept. of Computer Science, University of Rochester.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca A. Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *CoRR*, cs.CL/0006023.

Zachary O. Toups, William A. Hamilton, and Sultan A. Alharthi. 2016. Playing at planning: Game design patterns from disaster response practice. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '16, pages 362–375.

Christian Willms, Constantin Houy, Jana-Rebecca Rehse, Peter Fettke, and Ivana Kruijff-Korbayová. 2019. Team communication processing and process analytics for supporting robot-assisted emergency response. In *International Conference on Safety, Security, and Rescue Robotics (SSRR)*.

Fereshta Yazdani, Gayane Kazhoyan, Asil Kaan Bozcuoğlu, Andrei Haidu, Ferenc Bálint-Benczédi, Daniel Beßler, Mihai Pomarlan, and Michael Beetz. 2018. Cognition-enabled framework for mixed human-robot rescue teams. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 1421–1428. IEEE.

Appendix A

Team Communication Example

TL *Andreas, Andreas from Markus, come in.*

OP *yes, Andreas come in.*
$< \ldots >$

OP *yes, for information, I am ready* $[EHM]$ *shall I go ahead with my search command, or begin?*

TL *Yes, begin immediately without possible – least possible time delay, to* $[EHM]$ *have a higher chance for person rescue.*

OP *yes, understood, I begin with the search.*
$< \ldots >$

TL *Andreas from Markus, come in.* $[ent = unk.skippable]$

OP *Yes, Andreas, come in.*

TL $[ent = unk.skippable]$ *are there already any noteworthy findings?* $[ent = unk.skippable]$

OP *Negative. No noteworthy findings.* $[ent = unk.skippable]$

TL *Yes, understood.* $[ent = unk.skippable]$ *Daniel, Daniel from Markus, come in.* $[ent = unk.skippable]$ *Andreas from Markus, come in.*
$< \ldots >$

OP *Andreas, Markus from Andreas, come in.*

TL *Andreas, come in.*

OP *On first floor in the smoke found a barrel, green, labeled as environmentally hazardous material.*

TL *Yeah, can you* $[unintelligible]$ *whether anything is leaking?*

OP *Yeah. It is a 200 liter barrel, whether anything is leaking I cannot currently tell.*

TL $[EHM]$ *Any thermal emission?*

OP *No thermal emission.*

TL *Okay. Priority on continuing person search. Andreas from Markus, priority on continuing person search.*