Exploring Dialog Act Recognition in Open Domain Conversational Agents

Maliha Sultana[1] and Osmar R. Zaïane[1]

[1]University of Alberta, Alberta Machine Intelligence Institute, Canada.

Contributing authors: sultana2@ualberta.ca; zaiane@ualberta.ca;

Abstract

Recognizing dialog acts of users is an essential component in building successful conversational agents. In this work, we propose a dialog act (DA) classifier for two of our open domain dialog systems. For this, we first build a hierarchical taxonomy of 8 DAs suitable for classifying user utterances in open-domain setting. Next, we curate a high-quality, multi-domain dataset with over 24k user dialogs and annotate it with our 8 DAs. Next, we fine-tune our pretrained BERT-based DA classifier on this dataset. Through extensive experimentation, we show that our proposed model not only outperforms the baseline SVM classifier by achieving state-of-the-art accuracy but also generalizes extremely well on previously unseen data.

Keywords: Dialog Acts, Speech Act Recognition, Natural Language Processing

1 Introduction

Human beings are inherently social. Through frequent conversations, we convey our intentions, thoughts and opinions to our peers. Naturally, we grow accustomed to the everyday sentences we utter and the dialog acts we perform. In natural language understanding, a dialog act (DA) is an utterance in the context of a conversational dialog that serves a precise function in the dialog (or sometimes more than one) [1]. It can be a question, a statement or a request for action. Effective communication relies on recognizing the different DAs and responding accordingly. For example: someone asking a question

expects an answer as a response whereas someone giving an order expects its execution or acknowledgment of its execution.

Dialog systems have long been researched in the field of AI dating back to 1966 with the advent of Eliza, a chatbot [2]. Although intended to be a mere caricature of human conversation, users were soon treating ELIZA like a companion- confiding their most intimate thoughts. Nowadays, with the advancement in AI, chatbots are being used as virtual assistants in different fields to enhance productivity and reduce service costs. Recent studies have found that users often consider chatbots as friendly companions and not just mere assistants. In fact, over 40% of user requests received by customer service chatbots on social media have been observed to be emotional than informative [3]. How much trust a chatbot gains from its users depends on how human-like the chatbot is, i.e, how well it can handle natural language. As a result, recognizing the DA of users to generate better response has become an integral component in chatbots. Dialog systems usually include a taxonomy of dialog types or tags that are used to classify the different functions DAs can play. Depending on the task or domain in question, user intents vary and so do the proposed DA tag-sets. For example, to facilitate the development of dialog systems for mental-health counselling, Malhotra et al. [4] proposed a dataset called HOPE which consists of 12.9K patient-therapist utterances annotated with 12 dialog-act labels related to therapy. They also proposed a transformer based DA classifier which achieves state-of-the-art (SOTA) performance on HOPE.

With the aim of improving open-domain conversational agents, our work focuses on DA classification of users. We speculate that a dialog system can generate better responses through proper identification of user dialog acts. For this, we first identified the relevant dialog acts for our existing chatbots-MIRA[5] and ANA[6]. We then curated a corresponding high-quality, multi-domain dataset of ∼24k utterances belonging to one of our 8 proposed DAs-Statement, Factual Question, Yes/No Question, Direct Order, Indirect Order, Greeting, Feedback, Apology. Structuring this as a multi-class classification problem, we propose a pretrained BERT-base model as our DA classifier. Upon fine-tuning it on our curated dataset, the model achieves SOTA accuracy; outperforming the baseline SVM classifier by 3%. Our proposed DA classifier is also robust and generalizes well on never-before-seen dataset. In summary, the key contributions of our work are as follows:

1. We propose a hierarchical taxonomy consisting of 8 DAs suitable for open-domain conversational agents
2. We curate a high-quality, large-scale dataset of ∼24k user utterances from multiple domains and rich data sources
3. We propose a fine-tuned BERT-based model for DA classification which not only achieves SOTA performance on our dataset but also generalizes well on unseen data

The remainder of this paper is structured as follows: We present a summary of the related works in Sect. 2. We describe our proposed DA taxonomy in Sect. 3 and provide details on our data collection process. In Sect. 4, we present the architecture of our DA classifier and summarize the results of our comprehensive evaluation. Finally, we conclude our work in Sect. 6.

2 Related Works

Building conversational AI is a long-standing challenge in NLP. Human conversations are inherently complex and ambiguous. Training a dialog system that understands the semantic and syntactic nuances and generates natural and engaging response is difficult to achieve. However, recent works have shown the promise of combining dialog acts for neural response generation [7]. DAs can help conversational agents by providing a representation of the underlying meaning of a user's utterance.

In order to drive the research on building better dialog systems, a number of conversational corpora have been released in the past. The Switchboard Dialog Act Corpus (SwDA) [8] and the ICSI Meeting Recorder Dialog Act (MRDA) Corpus [9] are widely used to train dialog systems in open-domain setting. They consist of human-human utterances that are hand-labelled with over 40 dialog acts like Statement-non-opinion, Statement-opinion, Appreciation, Yes-No-Question, Wh-Question, Open-Question, Apology and so on. Authors like Colombo et al. [10] leveraged a sequence-to-sequence model and achieved an accuracy of 85% on SWDA, and SOTA accuracy of 91.6% on MRDA. Likewise, Li et al. [11] proposed a dual-attention hierarchical RNN with a CRF as their DA classifier. The model reached an accuracy of 92.2% on MRDA and SOTA accuracy of 82.3% on SWDA. On the other hand, Raheja et al. [12] proposed a DA classifier which can learn richer, more effective utterance representations with the help of self-attention and achieve an accuracy of 82.9% on SWDA and 91.1% on MRDA. More recently, to explicitly model the interaction between DA recognition and sentiment classification, Qin et al. [13] utilized co-interactive relation networks. Their classifier produced significant results for both tasks and even achieved performance boost after incorporating BERT [14]. Likewise, Saha et al. [15] jointly learnt dialog-act classification and emotion recognition tasks in a multi-modal setup.

Researchers have also looked into building DA classifiers for specific domains. To develop better learning environments and virtual mentors, Gautam et al.[16] proposed 8 unique dialog-act labels to classify their dataset consisting of student-mentor conversations in Nephrotex, a virtual internship. They also explored several machine learning methods to categorize the DAs and achieved promising results. Quinn et. al. [6] looked into improving their chatbot, ANA, by proposing 3 DAs: Declarative, Interrogative, and Imperative because they fit into ANA's definition of a potential user utterance. They used an SVM model as their DA classifier and achieved 72% accuracy on the dataset. Zhang et al. [17] proposed classifying Tweets into 5 user acts-

Statement, Question, Suggestion, Comment and Miscellaneous. Using a set of word-based and character-based features, their model achieved an average F1 score of nearly 0.70 on their dataset. Noticing how differently humans interact with other humans vs with machine, Yu et al. [18] proposed a DA annotation scheme called MIDAS based on human-machine conversations in open domain setting. The authors also collected and annotated 24k segmented sentences using MIDAS and deployed transfer learning to train a multi-label DA prediction model on it which achieved an F1-Score of 0.79.

Like the previous works, our paper aims towards building open-ended conversational agents that respond naturally by accurate detection of user DA. In particular, we focus on building a DA classifier that is applicable for our pre-existing text-based chatbots- ANA and MIRA. Apart from answering questions and sending reminders, Automated Nursing Agent or ANA aims to have a fluent and personalized conversation with the elderlies [6]. On the other hand, MIRA is a Mental Health Virtual Assistant which provides mental health resources to health care workers and their families [5]. It also has a module called 'Chatty MIRA' which allows users to have open-ended conversations with the chatbot. Given the difference in domain and task intents, our goal is to propose a common DA schema and its corresponding classifier. The next section gives a detailed explanation on how our DA tag-set was chosen.

3 Proposed Dialog Act Taxonomy

Although we had initially planned on curating a larger dataset using Quinn et al.'s [6] proposed dialog acts (Imperative, Declarative, Interrogative), we soon realised that their DA tag-set was too general and failed to capture multiple cases that require different response from the chatbot. For example: 'Can penguins fly?' and 'What is the name of our galaxy?' are both questions. However, the first one expects a yes/no answer whereas the second one expects a factual answer. To generate better responses, our chatbots need to learn how to distinguish between the two. After looking into the related works and having iterative discussions with two of our psychology students, we selected the following 8 DAs that adequately capture the intentions of our users. For a better understanding, Table 1 provides examples of each of the DAs categorized into a hierarchy.

1. Apology: Includes sentences through which the user expresses apology.
2. Greeting: Includes sentences through which the user greets the chatbot either at the beginning or towards the end of a session.
3. Informative: Includes queries asked by the user with the intention of gaining some information. Depending on the type of response, questions can broadly be of 2 types:
   (a) Yes/No: Includes close-ended queries that can be sufficiently answered with a simple yes or no.
   (b) Factual: Includes open-ended queries that seek fact-based answers. A majority of these questions are WH-questions but utterances like 'Name

Table 1: Selected dialog acts with examples

Dialog Act	Sub Category	Example
Apology		Sorry about that! My bad.
Greeting		Hey, how are you? Bye, see you soon!
Informative	Yes/No Factual	Is it possible to treat ADHD? What year did Bangladesh achieve their independence? Name the best therapist in my area.
Directive	Direct Order Indirect Order	Show me the list of hospitals nearby. I need help with managing anxiety. Can you turn on the music please?
Statement		I am being bullied at school lately. I like spending time with my family.
Feedback		This is exactly what I was looking for! Thanks. This is not what I wanted. You suck!

the highest rated therapist in my area' are also included here due to the similarity in user intent.

4. Directive: Includes orders given by the user to the chatbot for accomplishing a task. This again can be of two types:
   (a) Direct Order: Includes straightforward orders that are easy to detect, understand and carry out.
   (b) Indirect Order: Includes utterances that indirectly expect or request some type of action. These are a bit difficult to comprehend and might require the chatbot to first make an assumption and then prompt for a confirmation before execution. For example: 'I need help with managing anxiety' or 'Can you help me manage my anxiety?' can be interpreted as 'Show me resources for managing anxiety'.
5. Statement: Includes user utterances that do not request for an action or information. Rather, these are dialogs through which the user casually converses with the chatbot. By analyzing the emotion behind these utterances, the chatbot can either choose to give a sympathetic response or ask follow-up questions.
6. Feedback: Includes feedback from the user once the chatbot accomplishes a task e.g. carries out an order or answers a question. Feedback can be positive (when the chatbot is successful) or negative (when the chatbot is unsuccessful). By detecting the sentiment behind it, the chatbot can either thank the user or apologize and/or attempt the task again.

3.1 Data Sources

Once the DA taxonomy was decided upon, we moved onto curating the corresponding dataset. To make our chatbots can easily recognize user intents, we intended to include user utterances our conversational agents are likely to encounter. Moreover, since our goal is to build an open-domain DA classifier applicable to both of our chatbots, we want our training dataset to be versatile. For this, we included examples not only from mental health (for MIRA)

Table 2: Overview of our proposed training and test dataset

Dialog Act	Sub Category	Train Examples	Test Examples	% Distribution
Informative	Yes/No	3385	847	17.31
	Factual	3697	924	18.89
Directive	Direct Orders	3125	781	15.97
	Indirect Orders	5400	1349	27.60
Statement		3250	812	16.61
Greeting		239	60	1.22
Feedback		392	98	2
Apology		73	18	0.4

and popular chatbot commands (for ANA), but also from common domains like banking, air lines, product reviews and so on. We expect the mix of multiple data sources to add variation in sentence structure and make the dataset more diverse. For this, we first looked at some of the popular datasets that has a few dialog-act tags that are similar to ours. As for the rest, we scraped various websites and forums using simple rules. It is to be noted that, during data collection, we decided to include only those examples that followed our definition of each of the DAs. Moreover, to avoid dominance of a particular domain or type of sentence, we decided not to include too many examples from a single source. Below, we give a brief overview of the data sources we had used for each DA:

1. Informative: We used popular question-answering datasets and mental health FAQ websites.
   (a) Yes/No Question: BoolQ [19], SNIPS [20]
   (b) Factual Question: SNIPS[20], SQUAD [21]
2. Directive: We used task-completion dialog intent datasets to collect Direct and Indirect Orders. Simple extraction rules were used to distinguish between the two. Datasets include Taskmaster [22], SNIPS [20], ATIS [23] and ACID [24] to name a few.
3. Statement: We mostly used the dataset shared by a mental health forum called 'Counsel-Chat' which consists of anonymous user posts related to mental health. We also included some examples from Wiki-Article, IMDB Movie Review and Amazon Product Review datasets.
4. Feedback, Apology and Greeting: We scraped a few basic English learning websites to extract positive and negative appraisals, apologies and greetings.

Table 2 shows how the examples are distributed per class. Among the 8 dialog acts, Apology, Greeting and Feedback are our minority classes. Due to the lack of variation in the ways users greet, apologize and provide feedback in real life, these 3 classes have a small number of examples in comparison. In total, our dataset has 24450 examples. We split it into two in order to create a train and a test dataset. The split was done in a way to include 25% of the examples of each class into the test dataset in order to offset the class imbalance. Given that we used high-quality datasets as source, our curated dataset is also refined, with little to no mislabelling.

4 Proposed Dialog Act Classifier

Now, we will discuss in depth the architecture of our proposed DA classifier and report the results obtained through extensive experimentation. We further analyze the results and provide our inference.

4.1 Experimental Setup

Given the success rate of BERT in achieving SOTA result in multiple NLP tasks [25], our proposed DA classifier is a pretrained BERT-based model. BERT, which stands for Bidirectional Encoder Representations from Transformers is based on the transformer architecture that uses bidirectional training to have a deeper sense of language context. Moreover, because BERT was trained on a huge corpus, it can easily be fine-tuned on a new dataset and achieve great results. For the task of DA classification, we first convert our labels into categorical data. Next, we load the pretrained 'bert-base-cased' model from Tensorflow and fine-tune it on our training dataset. Next, we use the corresponding tokenizer with the maximum length set to 70. The BERT layers accept 3 input arrays but since 'tokenTypeIds' is necessary only for the question-answering model, we work with 2 input layers-'inputIds' and 'attentionMask'. We use also 'GlobalMaxPooling1D' and then a dense layer to build the CNN layers using hidden states of BERT. These CNN layers yield the output. We use the Adam optimizer with a learning rate of 5e-05, a decay of 0.01 and 'CategoricalCrossentropy' as loss. Once training is completed in 3 epochs, we calculate its accuracy on the test data.

To compare the performance of our DA classifier against a baseline, we use an SVM model. Support Vector Machine or SVM is a popular supervised learning algorithm that works by creating a decision boundary that best segregates an n-dimensional space into distinct classes. SVM is suitable for text categorization task as they can efficiently handle high dimensional input space, few irrelevant features and sparse document vectors [26]. Moreover, since text categorization problems are mostly linearly separable, SVM is the perfect candidate. They are very light-weight and are much faster to train than large language models. Given their benefits, a number of authors have successfully used SVMs for text classification tasks [6, 27–29]. We use LinearSVC as our baseline which is similar to SVC with the parameter kernel='linear', but provides more flexibility in the choice of penalties and loss functions in Scikit-learn [30]. As for converting the text files into numerical feature vectors, we use the Bag-of-Words (BoW) technique (CountVectorizer) and later run the TF-IDF technique (TfidfTransformer) over the features generated by BoW. Lastly, we train the baseline on our proposed dataset and evaluate it on our test dataset.

4.2 Performance Evaluation

From Table 3, it is evident that both the baseline and our DA classifier perform well on our proposed dataset. SVM yields an accuracy of 96% and BERT outperforms SVM by 3% by achieving an accuracy of 99%. One of the reasons

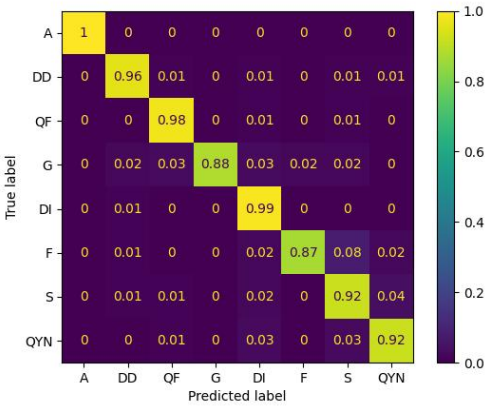
Table 3: Comparison of performance between the baseline (SVM) and our proposed classifier (BERT) on our proposed dataset

Model	Accuracy	Precision	Recall	F1-Score
SVM	0.96	0.96	0.94	0.95
BERT	0.99	0.99	0.99	0.99

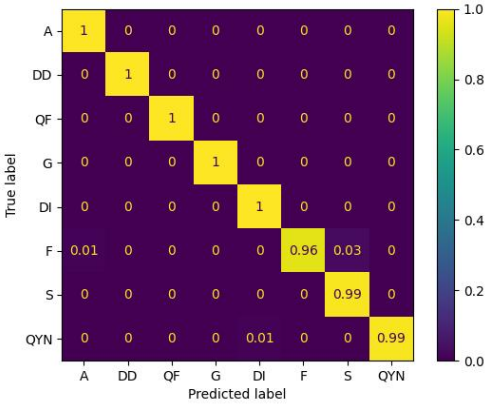
for such high accuracy rates might be because of the stark differences in the structure of sentences for each of the dialog acts. For further analysis of the wrongly predicted examples, we take a look at the confusion matrices in Figure 1 where the y-axis shows the true labels of the examples and the x-axis shows the predicted labels.

For majority classes (∼3771 examples/label) like Directive Direct Order (DD), Question Factual (QF) and Directive Indirect Order (DI), SVM achieves individual accuracies of 96%, 98% and 99% with only a very few instances of misclassification. The accuracy, however, is comparatively low for other majority classes like Statement (S) and Question Yes/No (QYN) (92%). Upon further analysis, we see that 8% of Feedback (F) are misclassified as Statement. For example: 'This works well ' and 'I'm glad you are my friend' are all Feedback but are wrongly predicted as Statement. This makes sense given the similarity in sentence structure for both of these classes. In case of Yes/No Question, the low accuracy rate comes from misclassifying a large number of Statement (4%) and Feedback (2%). For example: 'My issue is that there is always drama' and 'It is good' were wrongly predicted as Yes/No Question. Possible reason for this might be the presence of the helping verb 'is' towards the beginning of the sentence which is similar to the structure of a Yes/No Question ('Is it cold in here?'). The baseline model fails to learn the difference in these cases. As for the minority classes, although SVM scores a high accuracy for Apology (A) class, it struggles to detect Greeting (88%) and Feedback (87%) in comparison. Possible reason for this might be because the training data is not enough for the baseline to learn specific patterns to recognize them. Future work might look into using rules to detect these minority classes instead and compare the performance.

Now, we take a look at our fine-tuned pretrained BERT-based model. Unlike SVM, it does a very good job at achieving 99% accuracy for all seven of the eight classes. The Feedback class, however, has a slightly low accuracy rate (96%) for misclassifying some of the examples as Statement. Given the similarities shared by the examples of these two classes, it makes sense why the misclassification happened. On the bright side, it is important to mention that unlike the baseline, our BERT-based model does not struggle with accurately predicting Yes/No Question, Greeting or Statement. Thus, our proposed DA classifier not only outperforms the baseline model but also achieves SOTA result on our high-quality dataset.



(a) SVM



(b) BERT

Fig. 1: Confusion matrices of the baseline (SVM) and our proposed classifier (BERT) on our proposed dataset

4.3 Generalizability of Model

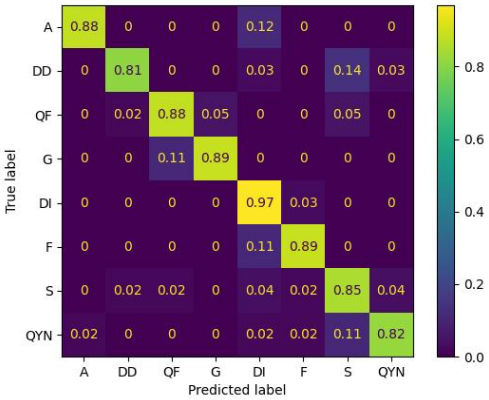
The generalizability (or robustness) of a model is a measure of its successful application to datasets other than the one used for training and testing. To compare and evaluate the generalizability of the baseline and our proposed DA classifier, we decided to create a new dataset called 'generalized dataset'. The plan was to find a data source that was never used for curating our original train and test data and then manually label it with our proposed taxonomy of 8 DAs. We chose the DialogSum dataset [31] for this purpose. It is a large-scale dialog summarization dataset consisting of ∼13k dialogs from

3 public dialog corpora, namely Dailydialog [32], DREAM [33] and MuTual [34], as well as an English speaking practice website. The dataset contains face-to-face high quality spoken dialogs from a wide range of daily-life topics including schooling, work, medication, travel and so on. Most of the conversations take place between friends, colleagues, and between service providers and customers. This, however, is an issue for us because we trained our model on a dataset that has conversations a user is more likely to have with a chatbot-not a person. We mitigated this by only including dialogs a user is more likely to have with a chatbot. For example: sentences like 'Zach, what's that on your arm?', 'Here, let me help you with your coat and we'll be on our way' were avoided. Moreover, given how large the DialogSum dataset is, we only chose a few samples for each DA manually. In the end, the curated generalized dataset consisted of 8 Apology, 9 Greeting, 9 Feedback, 30 Indirect Order, 36 Direct Order, 43 Factual Question, 45 Yes/No Question and 47 Statement.

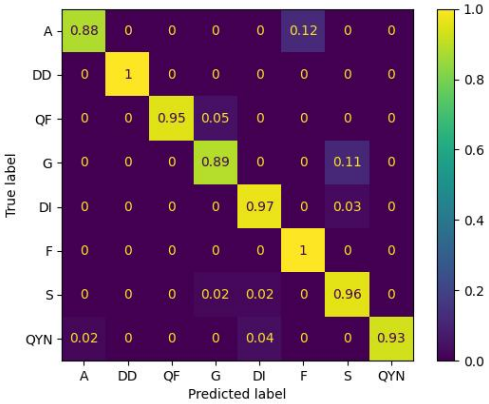
Now we evaluate the performance of the baseline and our proposed DA classifier on the generalized dataset. From Table 4, we can see that the performance of both the models drops which is expected. Given that the generalized dataset has more human-human conversations whereas our proposed dataset i.e the dataset the models were trained on has more human-machine conversations, this makes sense. However, what is impressive is that, although the accuracy of the baseline drops drastically by 10% on the generalized dataset (from 96% to 86%), our proposed DA classifier holds up really well. Even on the never-before-seen dataset, it achieves an accuracy of 96%- a mere 3% drop from its performance on our test dataset which is remarkable. This proves that our proposed DA classifier is both generalizable and robust on unseen data.

For further analysis, we take a look at the examples that were mislabelled. Figure 2 shows the confusion matrices for both the models. We can see that the baseline struggles the most with identifying Direct Orders. For example: sentences like 'Please wrap it for me and I'll take it', 'Go back to sleep then but only five more minutes' etc. are mislabelled as Statement and Yes/No Question. Since the model was trained on some very common chatbot commands like playing a song, booking a flight or reserving a seat- it has a hard time predicting these unconventional commands as Direct Order. The baseline also struggles with classifying a number of Yes/No Question correctly. For example: sentences like 'Excuse me, do you speak English?' and 'Have you turned on the air-conditioner?' are mislabelled as Statement and Indirect Order. This might be because our training dataset includes Yes/No Question that are usually factual and not casual (i.e sentences like 'Is Canada in the United States of America?' instead of 'Do you like to play the piano?'). Moreover, the phrase 'excuse me' in our dataset is mostly associated with the class Apology which might be another reason for the wrong prediction. On the flip side, the accuracy rate for the class Indirect Order is very high (97%). Overall, the accuracy rate of all the classes is above 80% which is not ideal but reasonable.

As for our proposed DA classifier, we notice the least accuracy in the minority classes Apology (88%) and Greeting (89%). This happens for mislabelling



(a) SVM



(b) BERT

Fig. 2: Confusion matrices of the baseline (SVM) and our proposed classifier (BERT) on generalized dataset

two sentences 'I hope you can forgive me' and 'Hi, my name is Susan' as Indirect Order and Statement. Probable reason for this is the lack of enough training data for these two classes. As a result, the model is not able to learn all sorts of variations properly. On the bright side, the remaining classes all have an impressive accuracy rate (over 90%). Despite having an accuracy of 93%, the Yes/No Question class struggles a bit with sentences like 'Have you turned on the air-conditioner', 'Can I exchange it?'. This might be because these sentences somewhat resemble the structure of Indirect Order examples present in the training data ('Turn on the air-conditioner', 'Exchange it'). All

Table 4: Comparison of performance between the baseline (SVM) and our proposed classifier (BERT) on the generalized dataset

Model	Dataset	Accuracy	Precision	Recall	F1-Score
SVM	Test	0.96	0.96	0.94	0.95
	Generalized	0.86	0.85	0.87	0.86
BERT	Test	0.99	0.99	0.99	0.99
	Generalized	0.96	0.92	0.95	0.93

in all, the experiments clearly prove that our proposed DA classifier generalizes far better than the baseline.

5 Conclusion

Classifying the intent of a user dialog in a conversation, also known as dialog act, is a key component in building conversational agents. By identifying the different DAs, chatbots can respond more coherently and assist users in accomplishing their tasks more effectively. In this work, we propose a BERT-based DA classifier for two of our open domain conversational agents- ANA and MIRA. For this, we first investigated the current literature and through iterative discussions, proposed a taxonomy of 8 DAs that are suitable for our chatbot users. We then curated a high-quality, large-scale dataset consisting of ∼24k user utterances from multiple domains. Upon fine-tuning our proposed classifier on this dataset, it outperforms the baseline SVM model by achieving SOTA accuracy. Through further evaluations, we prove the generalizability and robustness of our proposed model on unseen dataset. As for future work, we plan on investigating the effectiveness of structuring DA classification as a multi-label instead of a multi-class classification problem. We also want to look into the feasibility of including more dialog acts into our taxonomy.

References

[1] Popescu-Belis, A.: Abstracting a dialog act tagset for meeting processing. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04). European Language Resources Association (ELRA), Lisbon, Portugal (2004). http://www.lrec-conf.org/proceedings/lrec2004/pdf/268.pdf

[2] Wei, C., Yu, Z., Fong, S.: How to build a chatbot: chatbot framework and its capabilities. In: Proceedings of the 2018 10th International Conference on Machine Learning and Computing, pp. 369–373 (2018)

[3] Xu, A., Liu, Z., Guo, Y., Sinha, V., Akkiraju, R.: A new chatbot for customer service on social media. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 3506–3510 (2017)

[4] Malhotra, G., Waheed, A., Srivastava, A., Akhtar, M.S., Chakraborty, T.: Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, pp. 735–745 (2022)

[5] Noble, J.M., Zamani, A., Gharaat, M., Merrick, D., Maeda, N., Foster, A.L., Nikolaidis, I., Goud, R., Stroulia, E., Agyapong, V.I., *et al.*: Developing, implementing, and evaluating an artificial intelligence–guided mental health resource navigation chatbot for health care workers and their families during and following the covid-19 pandemic: Protocol for a cross-sectional study. JMIR Research Protocols **11**(7), 33717 (2022)

[6] Quinn, K., Zaiane, O.: Identifying questions & requests in conversation. In: Proceedings of the 2014 International C* Conference on Computer Science & Software Engineering, pp. 1–6 (2014)

[7] Welivita, A., Pu, P.: A taxonomy of empathetic response intents in human social conversations. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 4886–4899. International Committee on Computational Linguistics, Barcelona, Spain (Online) (2020). https://doi.org/10.18653/v1/2020.coling-main.429. https://aclanthology.org/2020.coling-main.429

[8] Godfrey, J.J., Holliman, E.: Switchboard-1 release 2. Linguistic Data Consortium, Philadelphia **926**, 927 (1997)

[9] Dhillon, R., Bhagat, S., Carvey, H., Shriberg, E.: Meeting recorder project: Dialog act labeling guide. Technical report, INTERNATIONAL COMPUTER SCIENCE INST BERKELEY CA (2004)

[10] Colombo, P., Chapuis, E., Manica, M., Vignon, E., Varni, G., Clavel, C.: Guiding attention in sequence-to-sequence models for dialogue act prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 7594–7601 (2020)

[11] Li, R., Lin, C., Collinson, M., Li, X., Chen, G.: A dual-attention hierarchical recurrent neural network for dialogue act classification. In: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), pp. 383–392 (2019)

[12] Raheja, V., Tetreault, J.: Dialogue act classification with context-aware self-attention. 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), pp. 3727–3733 (2019)

[13] Qin, L., Che, W., Li, Y., Ni, M., Liu, T.: Dcr-net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 8665–8672 (2020)

[14] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: 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), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423. https://aclanthology.org/N19-1423

[15] Saha, T., Gupta, D., Saha, S., Bhattacharyya, P.: Emotion aided dialogue act classification for task-independent conversations in a multi-modal framework. Cognitive Computation, 1–13 (2020)

[16] Gautam, D., Maharjan, N., Graesser, A.C., Rus, V.: Automated speech act categorization of chat utterances in virtual internships. In: EDM (2018)

[17] Zhang, R., Gao, D., Li, W.: What are tweeters doing: Recognizing speech acts in twitter. In: Analyzing Microtext (2011)

[18] Yu, D., Yu, Z.: Midas: A dialog act annotation scheme for open domain humanmachine spoken conversations. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 1103–1120 (2021)

[19] Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., Toutanova, K.: Boolq: Exploring the surprising difficulty of natural yes/no questions. 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), pp. 2924–2936 (2019)

[20] Coucke, A., Saade, A., Ball, A., Bluche, T., Caulier, A., Leroy, D., Doumouro, C., Gisselbrecht, T., Caltagirone, F., Lavril, T., et al.: Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. arXiv preprint arXiv:1805.10190 (2018)

[21] Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 2383–2392 (2016)

[22] Byrne, B., Krishnamoorthi, K., Sankar, C., Neelakantan, A., Goodrich, B., Duckworth, D., Yavuz, S., Dubey, A., Kim, K.-Y., Cedilnik, A.: Taskmaster-1: Toward a realistic and diverse dialog dataset. 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), pp. 4516–4525 (2019)

[23] Hemphill, C.T., Godfrey, J.J., Doddington, G.R.: The ATIS spoken language systems pilot corpus. In: Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990 (1990)

[24] Acharya, S., Fung, G.: Using optimal embeddings to learn new intents with few examples: An application in the insurance domain (2020)

[25] González-Carvajal, S., Garrido-Merchán, E.C.: Comparing bert against traditional machine learning text classification. arXiv preprint arXiv:2005.13012 (2020)

[26] Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: European Conference on Machine Learning, pp. 137–142 (1998). Springer

[27] Luo, X.: Efficient english text classification using selected machine learning techniques. Alexandria Engineering Journal **60**(3), 3401–3409 (2021). https://doi.org/10.1016/j.aej.2021.02.009

[28] Morales-Hernández, R.C., Becerra-Alonso, D., Vivas, E.R., Gutiérrez, J.: Comparison between svm and distilbert for multi-label text classification of scientific papers aligned with sustainable development goals. In: Advances in Computational Intelligence: 21st Mexican International Conference on Artificial Intelligence, MICAI 2022, Monterrey, Mexico, October 24–29, 2022, Proceedings, Part II, pp. 57–67 (2022). Springer

[29] Kambar, M.E.Z.N., Nahed, P., Cacho, J.R.F., Lee, G., Cummings, J., Taghva, K.: Clinical text classification of alzheimer's drugs' mechanism of action. In: Proceedings of Sixth International Congress on Information and Communication Technology: ICICT 2021, London, Volume 1, pp. 513–521 (2022). Springer

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

[31] Chen, Y., Liu, Y., Chen, L., Zhang, Y.: DialogSum: A real-life scenario

dialogue summarization dataset. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 5062–5074. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.findings-acl.449. https://aclanthology.org/2021.findings-acl.449

[32] Li, Y., Su, H., Shen, X., Li, W., Cao, Z., Niu, S.: DailyDialog: A manually labelled multi-turn dialogue dataset. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 986–995. Asian Federation of Natural Language Processing, Taipei, Taiwan (2017). https://aclanthology.org/I17-1099

[33] Sun, K., Yu, D., Chen, J., Yu, D., Choi, Y., Cardie, C.: DREAM: A challenge data set and models for dialogue-based reading comprehension. Transactions of the Association for Computational Linguistics **7**, 217–231 (2019). https://doi.org/10.1162/tacl_a_00264

[34] Cui, L., Wu, Y., Liu, S., Zhang, Y., Zhou, M.: MuTual: A dataset for multi-turn dialogue reasoning. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1406–1416. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.130. https://aclanthology.org/2020.acl-main.130