Conversational Agent for Daily Living Assessment Coaching

Aditya Gaydhani[1*], **Raymond Finzel**[2], **Sheena Dufresne**[3], **Maria Gini**[1] and **Serguei Pakhomov**[2]

[1]Department of Computer Science and Engineering, University of Minnesota
[2]Department of Pharmaceutical Care & Health Systems, University of Minnesota
[3]Department of Experimental and Clinical Pharmacology, University of Minnesota
{gaydh001, finze006, gahmx008, gini, pakh0002}@umn.edu

Abstract

We present preliminary work-in-progress results of a project focused on developing a conversational agent system to help with training certified assessors in conducting assessments of functioning in activities of daily living. To date, we have designed a modular task-based conversational agent system and collected hypothetical dialogue data required for training system components as well as a knowledge base needed to generate a wide variety of synthetic profiles of "individuals" being assessed. One of the key components of the system is the topic tracking module that determines the current topic of the conversation. We report the results of experiments with several machine learning approaches to topic/domain classification. The highest accuracy of 83% was achieved with a bidirectional long short-term memory (BiLSTM) model with pre-trained GloVe embeddings. In addition to these results, we also discuss some of the other challenges that we have encountered so far and potential solutions that we are currently pursuing.

1 Introduction

The use of Artificial Intelligence (AI) technology in the form of conversational agents (CA) has now expanded far beyond popular intelligent in-home assistants that are capable of answering basic questions about weather, trivia, driving directions, or music selection [Sciuto *et al.*, 2018]. For example, despite significant barriers to its adoption in healthcare, CA technology (mostly rule-based) is being actively investigated as a tool to assist patients and clinicians across multiple clinical contexts including diagnostic, prognostic, and treatment scenarios [Laranjo *et al.*, 2018]. Specific to the domain of functioning, the use of CA technology is also being investigated in the context of patient care and monitoring after the patient has been discharged from the hospital [Fadhil, 2018]. Assessment of functioning and functional status is a key target in multiple clinical contexts such as nursing, physical and occupational therapy, geriatric medicine, neurology, and rheumatology, among other health disciplines. It is also central to several non-clinical domains including disability and human services. One's ability to perform day-to-day activities independently relies on unimpaired cognitive, motor, and perceptual abilities. Significant impairment in these abilities typically results in a need for assistive devices or external supervision and/or assistance. In the United States, significant public resources are dedicated to providing assistance to those in need. In Minnesota, that assistance in allocated based on specific needs. The certified assessors perform assessments by conducting extensive face-to-face verbal interviews with the individuals referred for services and make recommendations for the level of support required to meet the person's needs. The interviews cover a broad range of areas including activities of daily living (ADLs: e.g., dressing, toileting, bathing, mobility, etc.) and instrumental activities of daily living (IADLs: e.g., preparing meals, managing finances, etc.). One of the desired goals of these assessments is to determine the degree of independence to which the person being assessed is able to preform ADLs and IADLs and to do so as consistently and uniformly as possible across multiple assessors. CA technology offers a potential for standardizing the training of certified assessors by simulating the interactions between assessors and persons being assessed in a uniform and reproducible fashion.

The long-term objective of our ongoing project is to develop a conversational agent system and infrastructure to support training of certified assessors in conducting the assessment of needs for social services. The purpose for developing a conversational agent is to a) assist in shifting the mode of conducting assessments from a questionnaire/survey style to a more free-form conversational/narrative style, and b) to standardize assessment outcomes across individual assessors. Towards this long-term objective, we have developed a prototype of the Conversational Agent for Daily Living Assessment Coaching (CADLAC) that relies on a database of historical assessments, conducted by Minnesota Department of Human Services, of ADLs and IADLS in order to generate synthetic profiles of individuals with varying levels of independence and needs. In this paper, we describe the high-level system architecture and its components, and report the results of experiments with machine learning approaches to maximizing the accuracy of the domain classification component. We also discuss the challenges encountered during the devel-

*Contact Author

opment of natural language understanding (NLU) and natural language generation (NLG) components and possible solutions with which we are currently experimenting.

2 Methodology

The high-level architecture of CADLAC system is shown in Figure 1. We followed the traditional modular CA system design [Ultes *et al.*, 2017] vs. an end-to-end design [Wen *et al.*, 2017] because the modular design is more suitable in the current early stage of the development when large amounts of training data needed for the end-to-end design are not yet available. Our modular design includes standard components such as the Topic Tracker (Domain Classifier), NLU and NLG modules, a Dialogue Manager consisting of the dialogue state tracking and policy components. In the current early stage of the project, we have been able to generate enough data to use machine learning in order to train some of the CA system components including the Topic Tracker and the NLU module designed to identify user intent and recognize named entities needed to match the input utterance/question to the database containing historical records from which we generated synthetic profiles to represent a variety of levels of functioning. The remaining components including the Dialogue policy are currently rule-based. This architecture is implemented using the open source MindMeld platform for conversational AI [1].

2.1 Data

We designed a survey to collect the data required to model the CA. The survey asked the assessors to recall some of their past assessments and provide hypothetical and anonymous examples based on verbal interactions they have had during those assessments focused on specific domains of functioning. The survey was administered to approximately 1,700 certified assessors. The resulting data consists of 2,900 short dialogues (up to 3 turns: see the example dialogue below) covering 18 domains within ADLs and IADLs: *Dressing*, *Grooming*, *Bathing*, *Toileting*, *Incontinence Management*, *Heavy Housekeeping*, *Light Housekeeping*, *Laundry*, *Financial Activities*, *Mobility*, *Transfers*, *Mode of Transfer*, *Positioning*, *Mode of Positioning*, *Food Consumption*, *Meal Preparation*, *Meal Planning*, *Fine Motor Skills*. Each turn consists of a question by the assessor and the response to that question provided by the person being assessed. Additionally, we collected characteristics of the person being assessed such as approximate age, gender, communication style (open vs. closed), and the degree of independence to which they are able to perform activities on the following scale: a) completely independent, b) requiring intermittent supervision, c) requiring supervision throughout the activity, d) requiring intermittent physical assistance, e) requiring physical assistance throughout the activity, f) completely dependent.

[1] https://www.mindmeld.com/

Example dialogue in the Dressing domain	
Assessor:	Tell me about how you get dressed after you are done in the bath.
Participant:	I can dress myself.
Assessor:	Including putting your shoes and socks on?
Participant:	It can be tough.
Assessor:	What about putting them on is hard?
Participant:	It is hard for me to bend that far. But I take it slow and I get it done. I sit on my lift chair while I do it.

Based on these hypothetical dialogues, we have defined and are currently continuing to refine an annotation schema to manually label semantic frames and their elements that may be useful for this application. For example, we defined the following frames for the *Heavy Housekeeping* domain using BRAT annotation schema format:

!Housekeeping_heavy	
vacuum	Place-Arg?:Home_location, Helper-Arg?:Person
scrub	Artifact-Arg?:Home_location, Place-Arg?:Home_location, Device-Arg?:Instrument, Helper-Arg?:Person
shovel	Place-Arg?:Home_location, Helper-Arg?:Person

An annotation of a short hypothetical dialogue using this schema focused on *Heavy Housekeeping* is shown in Figure 2.

We also collected de-identified historical assessment data for approximately 12,000 individuals. These data comprise a mix of structured and unstructured fields. The structured fields refer to the age, gender, communication style, and ability level of the person being assessed corresponding to the independence scale mentioned earlier. Unstructured fields capture free-text notes made by assessors during assessments consisting of brief descriptions of the challenges, preferences, and any assistive equipment for each ADLs and IADLs domain.

2.2 Synthetic Profiles

The CA is given a synthetic profile for every session of interaction. The synthetic profile gives a personality to the CA by defining its characteristics such as age, gender, communication style, and the degree of independence to which it can perform activities for all domains within ADLs and IADLs. The synthetic profile also holds information about the challenges, preferences, and assistive equipment used across all domains. The responses of the CA are based on the underlying synthetic profile.

The characteristics of the synthetic profile, particularly the independence levels, need to be consistent with each other. For example, a person who is unable to walk independently is most likely unable to do housekeeping independently. To ensure consistency, we use historical assessment data to gener-

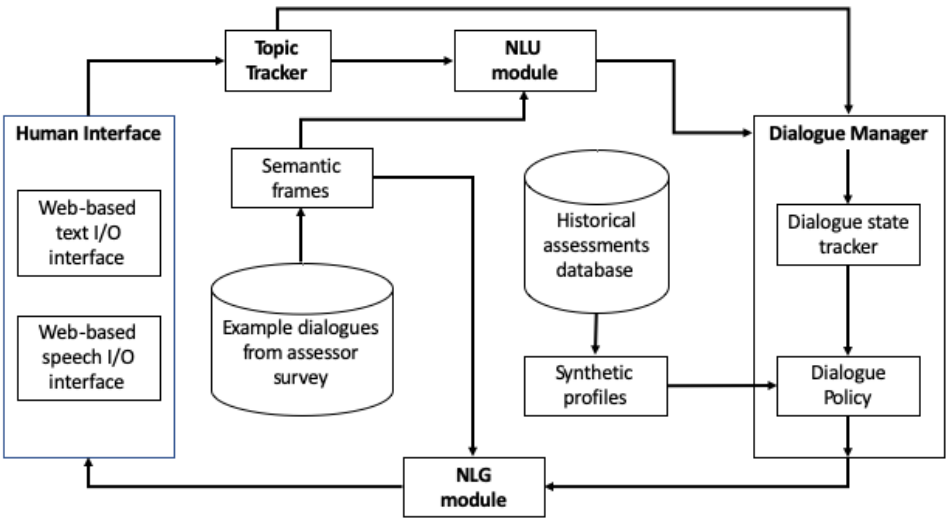

Figure 1: System architecture.

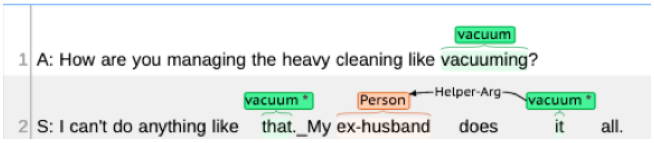

Figure 2: Example dialogue annotated for semantic frames.

ate synthetic profiles. At every session of interaction, a record is randomly sampled from the historical data and the fields of the synthetic profile are populated using this record.

2.3 Natural Language Understanding

The NLU module of the CA consists of domain classification, intent classification, and named entity recognition. The domain classifier or topic tracker determines the target domain for an input query. It performs a first-pass categorization of the incoming query and assigns it to one of the pre-defined domains. Each domain can have one or more intents that specify the task that the user wants to accomplish. The intent classifier identifies such intents for an input query. In this case, the input query is the question asked by the assessor to the synthetic profile of the CA. The question may consist of zero or more words or phrases, referred to as "entities", that need to be identified to generate an appropriate response. The named entity recognizer identifies such entities in the question.

One of the approaches to text classification is to use simple rule-based algorithms. These algorithms detect certain keywords in the incoming query and classify it into an appropriate class. However, such rule-based algorithms often have limited capabilities and do not generalize well. Moreover, the complexity of the rules increases with more variation in the type of input queries, hence these approaches are not scalable. In this paper, we explore more sophisticated machine learning and deep learning approaches to text classification.

2.4 Dialogue State Tracking

Conversational interaction consists of dialogue states, where each state is responsible for generating a particular type of response. Dialogue state tracking refers to mapping of incoming queries to appropriate dialogue states. We use an effective rule-based and pattern matching procedure in the CA for dialogue state tracking. The rules defined by this procedure rely on the domain, intent, or entities identified for an incoming query, as well as profile characteristics such as communication style. A dialogue state is determined by a combination of these attributes.

One of the challenges in modeling the CA is handling generic follow-up questions because such questions refer to the previous utterances of the conversation. We create a separate domain for generic follow-up questions using the assessor's questions from the 2nd and 3rd turn of the dialogues in the data. Whenever the system classifies an incoming query as a generic follow-up question, the domain of the previous turn is carried over to the current turn. Moreover, if the follow-up question does not consist of any entities of its own, then the entities from the previous turn are also carried over.

Communication style of the person being interviewed is one of the characteristics that we incorporate in the synthetic profile of the CA. Profiles with closed communication style are intended to generate brief responses that do not reveal details at the first utterance. It is important to track the questions corresponding to such utterances so that a detailed response can be generated after the assessor asks follow-up questions to the CA.

2.5 Natural Language Generation

The NLG module generates responses to the input queries. One of the common approaches used in NLG is delexicalization [Wen *et al.*, 2015], which is the process of using placeholders to represent slots in a sentence, which are then populated using the actual values of entities identified

from the input sentences. Recent studies [Xing *et al.*, 2017; Cai *et al.*, 2019] have also shown promising results using sequence-to-sequence models for dialogue generation.

One of the challenges in NLG for this application is that the responses are based on the identified attributes from the input query such as domain, intent, and entities, as well as the characteristics of the synthetic profile. Our current approach relies on using the unstructured text of the assessor notes contained in the historical database to generate responses to assessor questions that would match the topic and intent of the question and also would provide information consistent with the selected synthetic profile. For example, the first question in "Example dialogue for the Dressing domain" described above would be categorized as belonging to *Dressing* with the intent to elicit challenges that the person experiences in this domain. In this case, the question would be mapped to a specific synthetic profile in which the synthetic "person" is marked as *independent* in this ADL. The database entry for this profile would also contain assessor notes regarding challenges with dressing that may say "Able to dress on her own." The challenge for the NLG module is to "translate" this note into a natural language response such as "I can dress myself." In order to address this challenge we are currently experimenting with sequence-to-sequence machine translation modeling trained on manually generated data. This work is currently in progress.

3 Experiments

3.1 Domain Classification

Classification of text data using machine learning involves two tasks: transforming text into a numerical representation and feeding this representation into a classifier. We perform comparative analysis of various classification algorithms ranging from traditional machine learning approaches to modern neural networks for this task. We also explore techniques for extracting features from text.

Data Preparation. The dataset used to train the models was created from the data collected from the surveys. It comprises queries belonging to domains that fall under the categories of personal cares, household management, eating and meal preparation, and movement. We divided the conversation snippets from the surveys into turns and labeled them according to their domain. We also added data for small talk, in particular, a collection of phrases for greeting, interrogating and ending the conversation. We created a separate domain for generic follow-up questions. The resultant dataset consists of 20 domains and 2885 examples, and it is fairly balanced across the domains. 20% of this data was randomly sampled for testing and the remaining 80% was used for training the models.

Models. We included Logistic Regression, Support Vector Machines (SVM), Decision Trees, and Random Forests models as baselines. We tuned the hyperparameter settings of these models by performing an exhaustive grid search using 5-fold cross-validation. We compared the performance of these models with a Bidirectional Long Short-Term Memory (BiLSTM) neural network. LSTM [Hochreiter and Schmidhuber, 1997] is a type of Recurrent Neural Network (RNN) that has capabilities of learning long-term dependencies. It is widely used in sequential learning problems like language. The model architecture is shown in Figure 3. In the network, we used 20% spatial and recurrent dropout regularization [Srivastava *et al.*, 2014] to prevent overfitting. We set batch size to 64, and used ADAM [Kingma and Ba, 2015] optimizer and categorical cross-entropy loss.

Model	Acc.	F1-Score	F1-Weighted
LR	0.797	0.773	0.793
SVM	0.780	0.744	0.772
Decision Tree	0.706	0.670	0.700
Random Forest	0.710	0.669	0.699
BiLSTM	0.808	0.780	0.806
BiLSTM + GloVe	0.830	0.801	0.827

Table 1: Domain Classification Results

Feature Extraction. The baseline models use n-gram features that are extracted from the data corpus. In particular, we extract uni-gram, bi-gram, and tri-gram features. In recent years, distributed word representations [Mikolov *et al.*, 2013], or word embeddings, have shown impressive performance in various natural language processing tasks. In this paper, we make use of pre-trained GloVe embeddings [Pennington *et al.*, 2014] for our BiLSTM model. We also experiment with training the embeddings from scratch using the dataset.

Results. The results of the models are shown in Table 1. We use accuracy, f1-score, and weighted f1-score as our performance metrics for evaluation. The results show that the BiLSTM models outperform the traditional machine learning baseline models. Moreover, using pre-trained GloVe embeddings further improves the result of the BiLSTM model with embeddings trained from scratch. The BiLSTM model achieves 80.1% f1-score, 82.7% weighted f1-score, and 83% accuracy over a fairly balanced data. Analyzing the confusion matrix shows some level of misclassification among similar domains, e.g., planning meals and preparing meals, due to the similar nature of dialogues between these classes. Merging such domains increases the accuracy of this model to 94.2%.

4 Discussion

In this paper we presented some of the preliminary results of a work-in-progress project aimed at developing a conversational agent system for training certified assessors in conducting assessments for human services eligibility. The focus of the experiments reported here was on topic tracking for which we experimented with a range of machine learning approaches to text categorization. So far, we found that the best accuracy for domain categorization was achieved with a bidirectional LSTM model with pre-trained GloVe embeddings. Our modeling results also show that some of the distinctions between functional categories (e.g., *Positioning*, *Mobility*, *Transfers*, *Mode of Positioning*, and *Mode of Transfer*) are not supported by the currently available data and may require further data collection efforts in order to increase the accuracy of the topic tracker at a higher granularity.

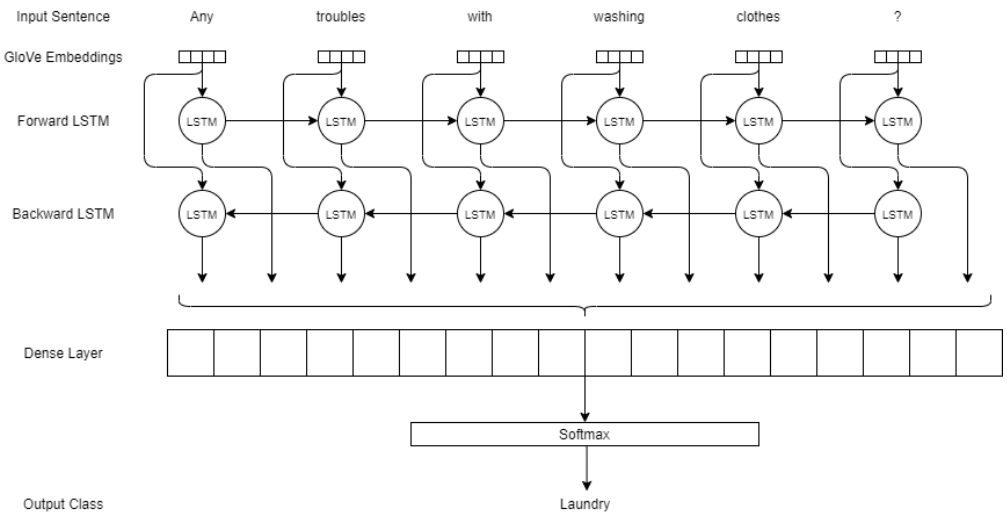

Figure 3: An illustration of the BiLSTM architecture

Our experiments with topic classification have a number of limitations. First, the data used for training and evaluation were collected as part of a survey in which assessors were asked to recall prior assessments, resulting in realistic but still hypothetical dialogues. The models developed on these data would need to be further evaluated on actual interviews between assessors and the persons being assessed, which is something we plan to do in future steps. Another potential limitation of the current CA system as a whole in the context of training certified assessors is that information gained by assessors through verbal interactions is only a part of what drives their assessments. Much of the additional information comes from non-verbal cues such as direct observation of the individual being assessed and the observation of the environment. Currently, our system is not designed as an embodied CA and does not provide non-verbal information about the physical environment in which the assessment is taking place.

Our next most immediate steps include training an intent classifier to recognize intents for all domains. Additionally, we intend to experiment with transformer based models to train a named entity recognizer to identify entities in the input queries, and use sequence-to-sequence models for the NLG component. We are also working on a strategy to provide feedback to the assessors regarding their conduct of the interviews and consistency of their assessments with synthetic profiles.

Acknowledgements

The work on this project was supported by funding from the Minnesota Department of Human Services. We would like to thank the people at DSD and MNIT for help with project specifications, gathering of historical data, and expert guidance on domain-specific aspects of the project.

References

[Cai *et al.*, 2019] Hengyi Cai, Hongshen Chen, Cheng Zhang, Yonghao Song, Xiaofang Zhao, and Dawei Yin. Adaptive parameterization for neural dialogue generation. 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 1793–1802, Hong Kong, China, November 2019. Association for Computational Linguistics.

[Fadhil, 2018] Ahmed Fadhil. Beyond patient monitoring: Conversational agents role in telemedicine and healthcare support for home-living elderly individuals. arXiv:1803.06000 [cs.CY], 2018.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. In *3rd International Conference for Learning Representations*, 2015.

[Laranjo *et al.*, 2018] Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Y S Lau, and Enrico Coiera. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258, July 2018.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, 2013.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. 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 2014. Association for Computational Linguistics.

[Sciuto *et al.*, 2018] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason Hong. "Hey Alexa, what's up?": A mixed-methods studies of in-home conversational agent usage. In *DIS '18: Proceedings of the 2018 Designing Interactive Systems Conference*, pages 857–868, June 2018.

[Srivastava *et al.*, 2014] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[Ultes *et al.*, 2017] Stefan Ultes, Lina M. Rojas Barahona, Pei-Hao Su, David Vandyke, Dongho Kim, Iñigo Casanueva, Paweł Budzianowski, Nikola Mrkšić, Tsung-Hsien Wen, Milica Gasic, and Steve Young. PyDial: A multi-domain statistical dialogue system toolkit. In *Proceedings of ACL 2017, System Demonstrations*, pages 73–78, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[Wen *et al.*, 2015] Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015.

[Wen *et al.*, 2017] Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain, April 2017. Association for Computational Linguistics.

[Xing *et al.*, 2017] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. Topic aware neural response generation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3351–3357. AAAI Press, 2017.