Intelligent Voice Agent and Service (iVAS) for Interactive and Multimodal Question and Answers

James Lockett[1], Sanith Wijesinghe[1], Jasper Phillips[2], Ian Gross[2], Michael Schoenfeld[1], Walter T. Hiranpat[1], Phillip J. Marlow[1], Matt Coarr[2], and Qian Hu[2(✉)]

[1] The MITRE Corporation, 7515 Colshire Drive Virginia, Bedford, USA
[2] The MITRE Corporation, 202 Burlington Road, Bedford, MA 01730, USA
qian@mitre.org

Abstract. This paper describes MITRE's Intelligent Voice Agent and Service (iVAS) research and prototype system that provides personalized answers to government customer service questions through intelligent and multimodal interactions with citizens. We report our novel approach to interpret a user's voice or text query through Natural Language Understanding combined with a Machine Learning model trained on domain-specific data and interactive conversations to disambiguate and confirm user intent. We also describe the integration of iVAS with voice or text chatbot interface.

Keywords: Intelligent and multimodal · Interactive question and answer · NLU · Machine Learning

1 Introduction

Although voice assistants and text-based chatbots are commonly deployed for commercial product and service use cases, their ability and flexibility to respond to natural language queries specific to government customer service questions are minimal and emerging (Herman 2017; Hendry 2019). This is due to the large variation and expressiveness of human language across a diverse citizen demographic and the lack of domain knowledge specific to government services. In this study, we describe our approach to address this challenge using combined speech technology and natural language processing together with robust Natural Language Understanding (NLU), Machine Learning (ML) and Artificial Intelligence (AI). We describe use cases that demonstrate automated response to user query, call classification, and call routing. In Sect. 2, we describe the Intelligent Voice Agent and Service's (iVAS) NLU and ML model applied to use cases and discuss its classification performance. In Sect. 3, we further describe the iVAS application prototype's multimodal and intelligent question and response features.

A. Cuzzocrea et al. (Eds.): FQAS 2019, LNAI 11529, pp. 396–402, 2019.
https://doi.org/10.1007/978-3-030-27629-4_36

2 Robust Natural Language Understanding and Machine Learning to Interpret Users' Questions and Deliver Relevant Responses

Automatic Speech Recognition (ASR) technologies continue to advance and can now provide single-digit word error rate for conversational speech (Saon et al. 2017, Xiong et al. 2016, Park et al. 2019) to transcribe a caller's speech. ASR allows callers to express their questions through natural human language instead of using specific keywords or pressing numbers on a phone. However, even 100% transcription accuracy and complete sentence construction is insufficient for a purely rule-based language processing application. Due to the variations and expressiveness of human language understanding, interpreting a user's query or request remains a challenging task (Hu et al. 2019). For a system to generate the most relevant answer to a question, additional information is needed to refine the user's query, and this often requires a dynamic interaction between the user and the system.

In this study we describe our approach to enabling automated call and query classification and determining an appropriate response with domain-adapted word embeddings (Sarma et al. 2018). The approach combines speech technology with natural language processing and NLU, ML, and domain-specific AI. We use transfer learning from a pre-trained word-embedding model to provide features robust to lexical variation and word order in the user's speech or text query. Language model pre-training has been shown to be effective for improving many natural language processing tasks (Dai and Le 2015, Peters et al. 2018, Radford et al. 2018, Howard and Ruder 2018).

In the iVAS system, users are prompted by a synthesized voice agent or chatbot to express their needs in natural language. The user's speech is captured and transcribed as text input to a trained classifier that outputs a probability for each of the possible responses that are used by our system to intelligently deliver or execute the response to the query. Unlike rule-based systems or keyword matching, which need to be maintained by domain experts and are sensitive to both lexical and syntactic variations, our approach is data-driven and robust to variation in the user's natural speech or text query. Thus, it understands the semantic content even when it is represented by different words such as synonyms and acronyms, sentence structures such as questions or requests, and filler words. It allows users to express their needs or queries through natural language instead of relying on prescribed phrases or keywords.

2.1 Methodology

The use cases explored in this study look to automate the call routing function when a user calls into a service line. The objective is to reduce the wait times involved with talking to a customer service representative. We investigate how to automatically interpret a caller's intent through natural language and route the caller to the

appropriate call center service department. For two separate use cases, we consider 7 and 10 different service department routing options. The automatic call interpretation and routing task is framed as an N-way classification given an arbitrary utterance of English text to determine the service department that has the highest probability of being associated with that utterance.

Ideally, there would be actual historic call routing data with either transcripts or recordings that detail what a caller said that led to be routed to a given service department. In the absence of historical call or text query data or transcripts, our classifier model is based on a pre-trained Glove model (Pennington et al. 2014) and trained using data scraped from various online sources, including: (1) review sites such as Yelp, and (2) community forums related to the domain of the call center. We extracted sentences from these sources and manually labeled them for each of the possible responses. Each sentence is preprocessed by standard NLP techniques, e.g., case normalization, stop word removal, and punctuation removal. A pre-trained word embedding model maps each word in a sentence to a 300-dimensional vector. The vectors are averaged to obtain a final averaged word embedding representing the entire sentence. We trained a logistic regression classifier on these sentence vectors with the route/specific phoneline as the target of prediction in the N-way classification.

2.2 Results

The study on 2 separate use cases yielded F1 scores that ranged from 84% to 97% and 56% to 94% respectively (using 5-fold cross validation) as shown in Tables 1 and 2. Natural language expression queries to the call center can be correctly classified for the user's intended destination of the call with an average of 91% correct classification rate for use case 1.

Table 1. Classification scores for use case 1 (7 service departments)

Department	Precision	Recall	F1-score
Human resource	0.98	0.97	0.97
Eyes	0.95	0.9	0.93
Dental	0.87	0.95	0.91
Laboratory	0.95	0.87	0.91
Pharmacy	0.9	0.92	0.91
Radiology	0.92	0.82	0.87
Mental health	0.87	0.81	0.84
Average			91%

Table 2. Classification scores for use case 2 (10 service departments).

Department	Precision	Recall	F1-score
Dental	0.86	0.95	0.90
Education benefits	0.87	0.94	0.90
Eyes	0.95	0.89	0.92
Human resource	0.80	0.81	0.80
Laboratory	0.95	0.88	0.91
Mental health	0.87	0.79	0.83
Burial service	1.00	0.89	0.94
Pharmacy	0.89	0.94	0.91
Radiology	0.95	0.80	0.87
Vocation & rehabilitation	0.65	0.49	0.56
Average	0.88	0.84	0.85

When the same ML model trained and tested for use case 1 is applied to use case 2, in which 3 new routing options are added, the classification rate drops to an average F-score of 85%. This may mean more training data is needed to extend the model. We suppose bidirectional and contextualized word embedding models (Howard and Ruder 2018; Peters et al. 2018; Radford et al. 2018) are needed for classification tasks with a larger set of class members.

2.3 Implication and Application

This data-driven, ML approach can be generalized to other classification and response domains to make the question and answer system more robust and flexible to a user's natural language queries. We expect our call classification and routing NLU model can be further refined when trained on actual historical call center data.

3 Intelligent, Interactive, and Multimodal iVAS Prototype

In this section, we describe the iVAS question and answer prototype that implements the underlying NLU and ML models, which enable intelligent natural language voice conversation or text-based dialogue to obtain relevant answers from the system. The iVAS system is designed to (1) interpret various permutations of natural language through semantic parsing, expansion, and context-sensitive NLU, (2) interact with users to obtain user-specific information to refine/confirm the question and generate more relevant answers based on application domain knowledge, and (3) perform multimodal interaction through speech and written text for question and answer as illustrated in Figs. 1 and 2 respectively.

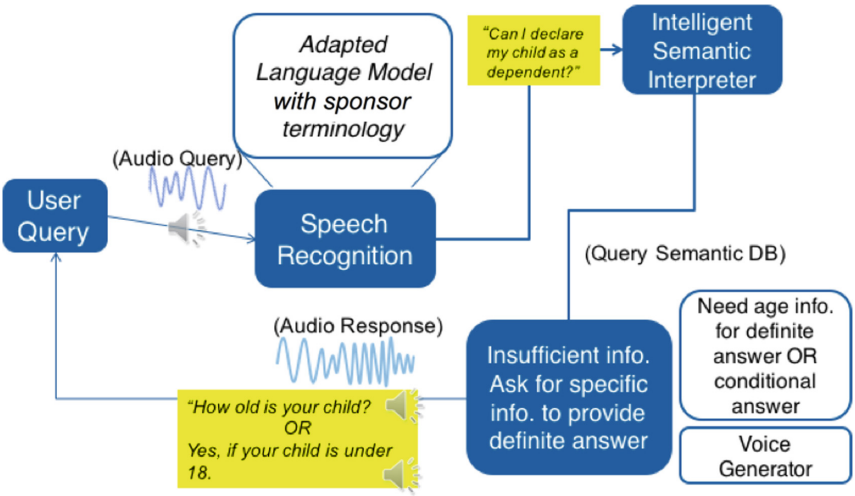

Fig. 1. Interactive natural language question and answer via speech

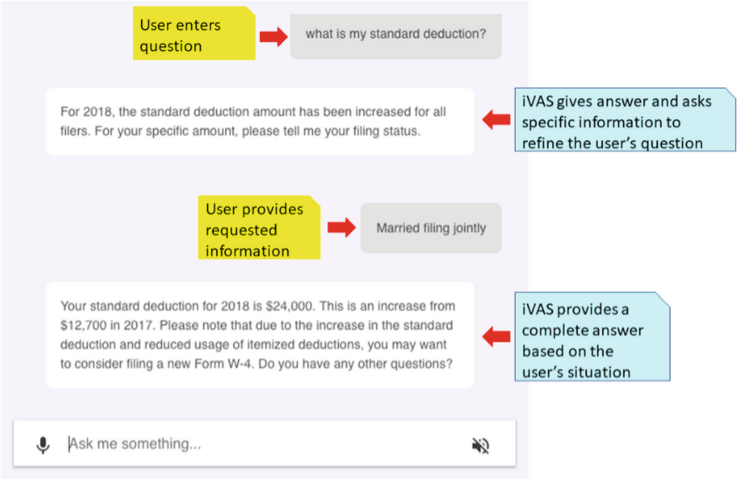

Fig. 2. Interactive natural language question and answer via chatbot

3.1 iVAS Prototype: Disambiguation via Interactive Conversation or Dialogue with the User

The iVAS voice and chatbot interface enables the conversation and dialogue between the user and the system. Even when the NLU and ML model finds a candidate response with a high system confidence, iVAS will still ask the user to confirm if the response is indeed what the user wants before executing the downstream process. This is illustrated in Fig. 3 in which iVAS finds the response with high probability.

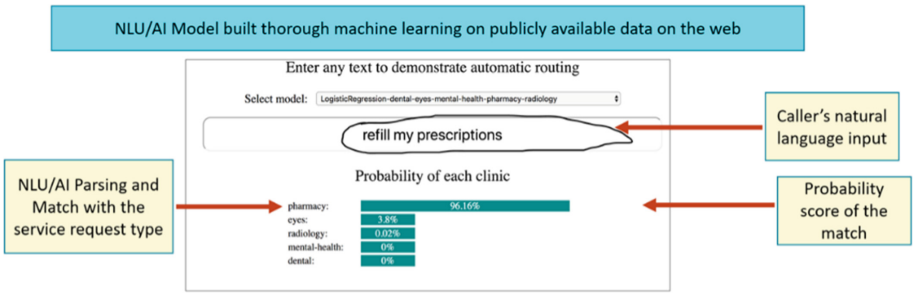

Fig. 3. iVAS interprets the user's query and identifies the service department with high probability.

Through natural language interaction via speech or text, the iVAS NLU and ML model automatically identifies the semantic underpinning of the user's intent or query and retrieves the response candidate from the system. If the system finds an answer or response to the question with a high system confidence, the answer or the response (such as routing to the matched service department) will be executed only after confirming with the user.

In cases when the iVAS NLU model finds two response candidates with close probability, the iVAS voice agent or chatbot asks the user to select the system-generated responses or make another query to ensure the system correctly interprets the user's intent or query.

3.2 iVAS Functional Blocks Supporting Additional Response Types

During the research and development of iVAS, it has become apparent that an intelligent system can provide other useful responses to a user's natural language query beyond just automatic call routing. The following is a list of the main response types the iVAS system can provide through interactive question and answer exchange: (1) automatically interpret/parse the intent of the incoming calls through NLU and a ML based domain-specific model, (2) interact with the caller/user when key information is needed to derive the most pertinent answer, (3) conduct automatic look-up and matching of the closest business facility/options based on the zip code or place name provided by the user, (4) verify/confirm the answer with the user before call routing, (5) automatically transfer the call to a live operator when urgent/crisis intent is identified/expressed, and (6) offer direct call transfer, provide a phone number or address, provide a relevant answer to the user's question or direct the user to a web portal to complete the business transaction.

3.3 iVAS Multimodal Architecture

iVAS is designed to be configuration driven and robust to voice and text question and answer interaction. Its architecture is agnostic of platform, device, and modalities (voice, text). Its NLU and decision-tree framework can be adapted to support various business and service needs for natural conversational query and answer systems.

4 Implication and Future Directions

The iVAS research and prototype demonstrates how, through natural conversation, relevant answers or responses to a user's query can be provided by speech or a text chatbot. There are larger classification tasks and more complex queries that require further investigation using efficient methodologies and deep learning of domain knowledge to provide the most relevant answers to a user. The research team is continuing the research of robust natural language query and answer system to provide efficient and useful answers to the users.

References

Herman, J.: U.S. federal AI virtual assistant pilot. In: Presentations Made at the U.S. General Services Administration Emerging Citizen Technology Program, Washington, D.C, 17 May 2017

Hendry, J.: DHS attempts most ambitious Microsoft digital assistant build yet, ITNews, 29 March 2019. https://www.itnews.com.au/news/dhs-attempts-most-ambitious-microsoft-digital-assistant-build-yet-523101

Saon, G., et al.: English conversational telephone speech recognition by humans and machines, [cs.CL], 6 March 2017. arXiv:1703.02136v1

Xiong, W., et al.: Achieving human parity in conversational speech recognition (2016). arXiv:1610.05256

Park, D.S., et al.: Specaugment: a simple data augmentation method for automatic speech recognition, 18 April 2019. arXiv:1904.08779

Dai, A., Le, Q.: Semi-supervised sequence learning. In: Advances in Neural Information Processing (NIPS) (2015)

Hu, Q., Lockett, J., et al.: Automated call classification and routing with speech technology and artificial intelligence. In: SpeechTek Conference, Washington DC, USA (2019)

Sarma, P.K., Liang, Y., Sethares, W.A.: Domain adapted word embeddings for improved sentiment classification. In Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP, pp. 51–59, Melbourne, Australia, 19 July 2018. c 2018 Association for Computational Linguistics

Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, pp. 1532–1543, 25-29 October 2014. c 2014 Association for Computational Linguistics

Peters, M., et al.: Deep contextualized word representations. In: NAACL (2018)

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding with unsupervised learning, Technical report (2018). OpenAI

Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: ACL. Association for Computational Linguistics (2018)