Modeling User Utterances as Intents in an Audiological Design Space

Benjamin Johansen
Jakob Eg Larsen
benjoh@dtu.dk
jaeg@dtu.dk
DTU Compute
Technical University of Denmark
DK-2800 Kgs.Lyngby, Denmark

Michael Kai Petersen
Niels Henrik Pontoppidan
mkpe@eriksholm.com
npon@eriksholm.com
Eriksholm Research Centre
Rortangvej 20
DK-3070 Snekkersten, Denmark

ABSTRACT

The global number of people living with hearing loss continues to grow, while the clinical resources are limited. To address this we describe a scalable goal oriented system. We outline a method on creating an audiological vocabulary, which can be mapped to intents. We create a shared audiological parameter space, with inspiration from clinical workflows. Matching of the intents and the audiological space, results in hearing aid fitting parameters, which then receive feedback from the user. We discuss how to train embedding and recurrent neural network models implementing attention mechanisms, to predict the optimal settings based on learned sequences of dialogue states and device fitting outcomes.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **HCI theory, concepts and models**;

CHI '19, May 05, 2019, Glasgow, UK

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

ACM Reference Format:
Benjamin Johansen, Jakob Eg Larsen, Michael Kai Petersen, and Niels Henrik Pontoppidan. 2019. Modeling User Utterances as Intents in an Audiological Design Space . In *Proceedings of CHI '19: Workshop on Computational Modeling in Human-Computer Interaction (CHI '19).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/1122445.1122456

INTRODUCTION

Computational interaction is an emerging research field spanning from optimizing input and interaction techniques using control theory and Hidden Markov Models. An emerging field within computational interaction is goal oriented interaction. An example of this is conversational interfaces. Notably the Google Duplex AI system capable of carrying out natural phone conversation to reserve a table at a restaurant, or make an appointment with a hairdresser [5]. The primary challenges of conversational interfaces are; the interface lacks understandable boundaries, the interaction mimics human behavior and is expected to act accordingly, the interface needs to have a robust speech-to-text engine, and the interfaces needs attention or memory to stay focused on the dialogue. Often, one or more of these challenges are not addressed, and triggers a mismatch between user expectations and interface performance. State of the art conversational systems overcome these challenges by thoroughly mapping out design boundaries, rather than attempting to encompass a full conversational dialogue. As an example, humans are capable of adapting dynamically to a changing context, and still revert back to the initial topic. We are able to do this through memory and attention mechanisms, and we understand the unvoiced boundaries of the conversation.
In the present paper we discuss how to enable interactions in an audiological design space, by embedding and training recurrent neural network models with simple attention and memory components. The goal is to predict the optimal hearing aid settings in real life listening scenarios.

AUDIOLOGICAL DESIGN SPACE

We focus on the use case of conversational agents within hearing health care, and on hearing aid fitting and optimization. The current clinical work flow is sequential, relies on calendars, and experienced hearing care professionals. The main challenge is lack of scalability. This is evident in emerging markets such as China and in low income countries, where the later has less than 1 audiologist per million citizen [13]. General health and medical care are facing similar challenges, where the number of patients are growing faster than the number of health care practitioners. Combining mobile internet connectivity with conversational interfaces may enable us to provide scalable healthcare solutions.

Voice enabled digital assistants, implementing artificial intelligence, are rapidly changing how we interact with internet of things devices including car dashboards and smartphone connected hearing

aids. The most successful goal-oriented dialogue systems, model conversations as partially observable Markov decision process (POMPDS) [15]. However, these goal-oriented application requires a lot of domain-specific handcrafting of features, which restrict their usage to specific domains. This hinders scalability and transfer learning to new domains [1]. The lack of annotated vocabulary for the audiological design space limits the use of POMPDS. It requires extensive effort to collect and annotate dialogues related to audiological trouble shooting. However, know-how of clinical practices can help establish a framework and context for dialogues. We draw inspiration from several studies where hearing care professionals map utterances into hearing aid fitting parameters [2, 11]. These parameters are related to frequency specific gain, loudness perception, and thresholds for attenuation and noise reduction. Based on the clinical practices, we outline an audiological design space.

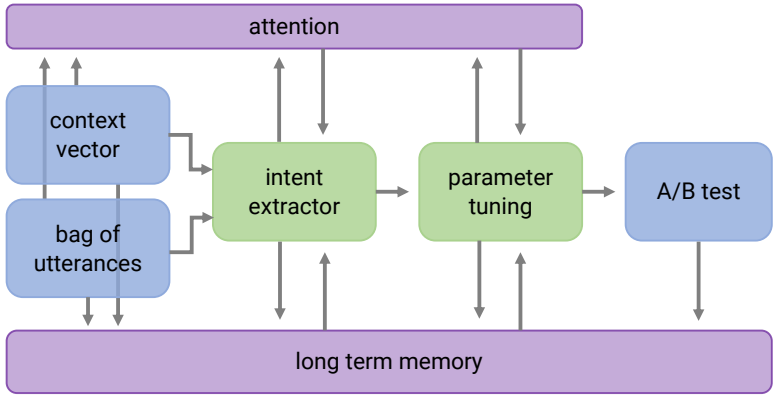

Figure 1: The proposed conversational agent uses both contextual and user input. It then probabilistic proposes a A/B program pair. The user picks a preference, and the memory network is updated.

We propose an interactive conversational agent, based on attention mechanisms mimicking human memory. The agent uses a context matrix and a utterance matrix as input, which is then fed through an intent extractor. The model includes both an attention unit for short term memory, using the current inputs to update parameters, and an attention network utilizing previous learned weights. Intents are not only inferred from semantics but also include comparison of four contrasting programs representing audiological parameters. The user is presented with an A/B program pair to select the preferred hearing aid setting. The memory is continuously updated based on a recurrent neural network model. An overview of the conversational agent is presented in Figure 1

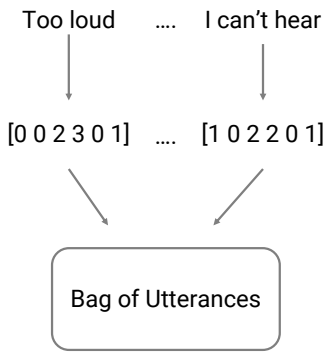

Figure 2: Using utterances to create a bag of utterances, a vocalbury. Our model use cosine similarity between utterances. The similarity is later used for audiological parameters.

MODELING UTTERANCES AS INTENTS

Word embeddings have been demonstrated to be an effective procedure in natural language understanding, as demonstrated by Mikolov et al.[6, 7]. The concept of skip-grams can be applied to longer sequences such as sentencens, or on documents, to create sequence embeddings [9].
We use the same principles of word embeddings to train a natural language understanding part of our model. We start by creating a vocabulary based on user utterances. Due to the model flexibility, we can create a new vocabulary from scratch. As an example, *'There is too much noise'* and *'I can't hear because there are too many people'*, have embedding vectors more similar than either *'Turn down the volume'* or *'It's too quiet'*, this is illustrated in Figure 2.

The model infers the most likely labels of new utterances, based on their cosine similarity to previously learned word vector representations. The feature and intents vectors have the same dimensionality, allowing the model to be trained by simply maximizing the cosine similarity between utterances and fitting label embeddings. A cosine similarity matrix of a selection of utterances are illustrated in Figure 3

We use a similar approach to embed intents from utterances. For example, *'I cannot hear the speaker in front of me'*, can relate to intents of focusing on the person, reducing surrounding noise, increasing volume output, or a combination of these. We cast the embeddings of utterances and intents into a shared low dimensional space using a supervised learning approach similar to StarSpace [14], implemented as a TensorFlow embedding model in Rasa [8, 10].

FROM INTENTS TO FITTING PARAMETERS

The first part of our model infers intents based on utterances from the user. The second part of the model search for optimal fitting parameters. The model fitting is based on empirical evidence on troubleshooting work-flows from hearing care professionals. A challenge within hearing care is the lack of *one to one* mapping between utterances and audiological solutions [2]. Meaning, a hearing care professional has to deduct a suitable hearing aid setting, by interpreting the challenges associated with the listening scenario the user has experienced. The fitting parameter labels thus resemble a flow chart of potential interventions [11].This requires the audiologist to interpretate utterances like *"it is very noisy"* which dependents on the listening scenario, the hearing loss compensation, and the cognitive state of the user. The audiologist has to estimate what the optimal audiological solution would be in a specific context. These solutions could involve highly different fitting parameters related to beamforming, noise reduction, loudness sensitivity or gain adjustments

To model such a clinical workflow, our goal oriented dialogue system needs to learn sequences of perceived intents, fitting actions and estimate the updated settings. Similar to the previous mapping of utterance to intents, we apply a supervised learning approach to train an embedding model. We

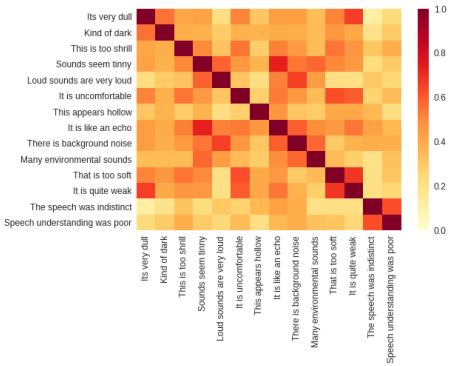

Figure 3: Confusion matrix of user descriptions of challenging listening scenarios based on semantic similarity generated by Universal Sentence Encoder; pairwise groups along the diagonal reflect how utterances are often mapped by audiologists to parameters of frequency specific gain, beamforming, occlusion and loudness.

then use a Recurrent Neural Network (RNN) to create sequential embeddings of perceived intents and fitting actions. This generates a memory state from previous dialogues and outcomes. Embedding dialogue states and fitting actions in the same vector space allows for comparing a new dialogue input against the system long term memory. Subsequently predicting the most likely audiological solution based on its cosine similarity to previously learned outcomes. Target labels or fitting actions can be represented as a bag of multiple features. Attention mechanisms enables the model to infer which intents and system actions contributed the most to previous outcomes, as shown by Vlasov et al. [12]. That is, separate user and system attention probabilities are inferred from the embeddings of past user inputs and fitting actions in order to adjust the weights determining the next hearing aid adjustment.

SIMULATING FITTING OUTCOMES BASED ON DIALOGUE STATES

Rather than relying solely on utterances for intent classification, we are able to generate contrasting hearing aid settings, which similar to AB testing enables the user to compare fitting parameters in real life listening scenarios [3]. Configuring four contrasting hearing aid settings, reflecting the fitting parameters that hearing care professionals commonly associate with utterances, enables users to indirectly accept or decline the suggested fitting parameters in a given context. Thus extending the underlying supervised learning of the goal oriented dialogue system with an additional reinforcement learning framework, to facilitate prediction of the preferred hearing aid settings. Defining the design space, according to the hearing aid adjustment parameters identified in previous studies [1, 11], we train an embedding model, in order to map utterances into intents related to adjustment of: maximum power output, frequency specific gain, beamforming, noise reduction thresholds, and loudness recruitment. We train the model based on multiple sequences of utterances, intents, hearing aid adjustments and final outcome. As exemplified in this simulated fitting dialogue we simply sequentially map utterances to intents:

utterance: "can't understand what is being said"
intent: enhance speech
action: increase B high frequency and soft gain
reply: "maybe try setting B ?"
utterance: "what does B do ?"
intent: explain B
reply: "it adds brightness to better separate sounds, do you prefer setting B ?"
utterance: "its very noisy in here"
intent: noise reduction
action: adjust B beamformer, attenuation and noise reduction thresholds
reply: "try setting B, do you prefer B ?"

utterance: "its better"
intent: confirm
reply: "great i'll remember you like setting B in this listening environment"

The embedding model is trained with multiple utterances for each of the intents to infer the most likely fitting parameters. These are used as dialogue flow inputs to train a RNN model, in order to predict fitting outcomes. As can be seen in the above simulated dialogue, utterances like *"can't understand what is being said"* might initially be mapped to the intent of enhancing speech based on gain fitting parameters. The subsequent utterance *"its very noisy in here"* shifts the intent towards adjusting beamformer, attenuation and and noise reduction thresholds.
Training on multiple dialogues, the TensorFlow RNN model applies attention mechanisms to learn which intents in a sequence contributed the most in order to predict to the final fitting outcome. These fitting parameters can furthermore be contextualized as the goal oriented dialogue system has access to continuous time series data describing the corresponding listening environment data [4]. Meaning, that the reinforcement learning of intents based on dialogues and fitting outcome, can be complemented with soundscape data, in order to automatically adjust hearing aid settings in real life listening scenarios.

FUTURE OUTLOOK

We suggest the following to be considered when designing flexible computational interfaces based on natural language understanding. 1) embeddings are useful for both understanding language, and for projecting other parameters into embeddings. This creates a shared embedding space, where different entities can be compared. 2) using attention mechanisms facilitates limiting the solution space. Learning from previous dialogue states and actions, helps the model to generate repsonses and predict the most likely next action. 3) our approach shows how to translate observed clinical workflows into parameter settings. This could be extended to general healthcare while supporting healthcare staff in the decision making process. 4) utilizing flexible and dynamic frameworks, such as the one we propose, continuously learn from interactions. This type of model can initially be trained on a small labeled data set, and continue to learn in a semi-supervised manner through user interactions.

ACKNOWLEDGMENTS

This work is supported by the Technical University of Denmark, Copenhagen Center for Health Technology (CACHET) and the Oticon Foundation. Oticon EVOTION hearing aids, and Niels Pontoppidan, are partly funded by European Union's Horizon 2020 research and innovation program under Grant Agreement 727521 EVOTION. We would like to thank Eriksholm Research Centre and Oticon A/S.

REFERENCES

[1] Antoine Bordes, Y-lan Boureau, and Jason Weston. 2016. Learning End-to-End Goal-Oriented Dialog. (may 2016), 1–15. arXiv:1605.07683 http://arxiv.org/abs/1605.07683

[2] Lorienne M. Jenstad, Dianne Van Tasell, and Chiquita Ewert. 2003. Hearing Aid Troubleshooting Based on Patients' Descriptions. *Journal of the American Academy of Audiology* 14, 7 (2003), 347–360.

[3] Benjamin Johansen, Maciej Jan Korzepa, Michael Kai Petersen, Niels Henrik Pontoppidan, and Jakob Eg Larsen. 2018. Inferring User Intents from Motion in Hearing Healthcare. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers - UbiComp '18*. ACM Press, New York, New York, USA, 670–675. https://doi.org/10.1145/3267305.3267683

[4] Benjamin Johansen and Michael Kai Petersen. 2018. Mapping auditory percepts into visual interfaces for hearing impaired users. In *Proceedings of the 2018 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, Montreal, 6.

[5] Yaniv Leviathan and Yossi Matia. 2018. Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone. https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html

[6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in neural information processing systems*. 3111–3119.

[7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. (jan 2013), 12 pages. arXiv:1301.3781 http://arxiv.org/abs/1301.3781

[8] Alan Nichol. 2018. Supervised Word Vectors from Scratch in Rasa NLU. https://medium.com/rasa-blog/supervised-word-vectors-from-scratch-in-rasa-nlu-6daf794efcd8

[9] Le Quoc and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Advances in neural information processing systems*. 1188–1196.

[10] RasaHQ. 2019. RASA. https://github.com/RasaHQ

[11] Thijs Thielemans, Donné Pans, Michelene Chenault, and Lucien Anteunis. 2017. Hearing aid fine-tuning based on Dutch descriptions. *International Journal of Audiology* 56, 7 (jul 2017), 507–515. https://doi.org/10.1080/14992027.2017.1288302

[12] Vladimir Vlasov, Akela Drissner-Schmid, and Alan Nichol. 2018. Few-Shot Generalization Across Dialogue Tasks. Nips (nov 2018). https://doi.org/arXiv:1811.11707v1 arXiv:1811.11707

[13] World Health Organization. 2013. *Multi-country assessment of national capacity to provide hearing care*. Technical Report. https://www.who.int/pbd/publications/WHOReportHearingCare{_}Englishweb.pdf

[14] Ledell Wu, Adam Fisch, Sumit Chopra, Keith Adams, Antoine Bordes, and Jason Weston. 2016. StarSpace : Embed All The Things ! (2016), 5569–5577.

[15] By Steve Young, Milica Gas, Blaise Thomson, and Jason D Williams. 2013. POMDP-Based Statistical Spoken Dialog Systems : A Review. 101, 5 (2013). https://doi.org/10.1109/JPROC.2012.2225812