Research Challenges in Building a Voice-based Artificial Personal Shopper - Position Paper

Nut Limsopatham
Amazon Research
Seattle
Washington, USA
nutli@amazon.com

Oleg Rokhlenko
Amazon Research
Seattle
Washington, USA
olegro@amazon.com

David Carmel
Amazon Research
Matam Park
Haifa, Israel
david.carmel@gmail.com

Abstract

Recent advances in automatic speech recognition lead toward enabling a voice conversation between a human user and an intelligent virtual assistant. This provides a potential foundation for developing artificial personal shoppers for e-commerce websites, such as Alibaba, Amazon, and eBay. Personal shoppers are valuable to the on-line shops as they enhance user engagement and trust by promptly dealing with customers' questions and concerns. Developing an artificial personal shopper requires the agent to leverage knowledge about the customer and products, while interacting with the customer in a human-like conversation. In this position paper, we motivate and describe *the artificial personal shopper task*, and then address a research agenda for this task by adapting and advancing existing information retrieval and natural language processing technologies.

1 Introduction

An intelligent virtual assistant, such as Amazon's Alexa, Microsoft's Cortana, Google Assistant and Apple's Siri, is essentially a voice-based dialog system that provides assistance to users for their daily activities, e.g. making a phone call, checking weather forecast, setting a reminder, and searching for relevant information (Thomas et al., 2017; Burtsev et al., 2017).

Several attempts have been made to develop an intelligent dialog agent using different approaches, including rule-based approaches (e.g. Weizenbaum 1966), machine translation (e.g. Ritter et al. 2011), information retrieval (e.g. Ji et al. 2014), classification (e.g. Shriberg et al. 1998), sequence-to-sequence models (e.g. Vinyals and Le 2015), reinforcement learning (e.g. Williams and Young 2007) and hybrid approaches (e.g. Bordes et al. 2016). Specifically, in his pioneering work, Weizenbaum (1966) developed the Eliza chatbot agent for interacting with patients with mental illness using syntactic rules. Ritter et al. (2011) learned to respond using phrase-based machine translation from Twitter conversations. Ji et al. (2014) learned to chit-chat from pairs of posts and an associated comments extracted from the Weibo social media platform. Vinyals and Le (2015) created an IT helpdesk dialog system using an encoder-decoder architecture based on recurrent neural networks. Their model converses by predicting the next sentence given the previous sentence or sentences in a conversation. Williams and Young (2007) modeled a dialog conversation as a partially observable Markov decision process (POMDP) and used reinforcement learning to optimize a response action at each time-step by maximizing the cumulative long-term reward.

In this position paper, we introduce a novel artificial personal shopper task, where a voice-based dialog system is used to enrich on-line shopping experience by replicating a personal shopping agent in a brick-and-mortar store. In particular, an effective artificial personal shopper would be able to converse and provide supports for the customer with information related to any products in the on-line store. Importantly, the assistance has to be personalized to individual customers. For example, in order to correctly answer the question "Is the Bose headphone compatible with my phone?", an artificial personal shopper has to know (1) what type of phone the customer has or refers to, (2) what is the model of the 'Bose headphone', and (3) whether *the headphone* is compatible with *the customer phone*. Table 1

The work describes the authors' ideas about the artificial personal shopper task and not of Amazon.

shows some examples of typical shopping related questions associated with potential responses.

The remainder of this paper is organized as follows. In Section 2, we define the artificial personal shopper task and discuss main information types required for handling this task. Section 3 describes some of the research challenges raised by the artificial personal shopper task, and how existing information retrieval (IR) and natural language processing (NLP) approaches could be applied for the task. Section 4 provides concluding remarks.

2 Types of Information Needed for the Personal Shopper Task

In the artificial personal shopper task, an intelligent virtual assistant provides personal shopping services by conducting a meaningful conversation with the customers. To achieve this task, we postulate that the personal assistant should be able to access and leverage three main types of information:

- **Product Information**: Information about the products is crucial for providing useful product-related conversations with the users. An artificial personal shopper should have an efficient and effective access to different forms of information related to each of the products that are available in the e-commerce store in order to answer factual questions about product attributes, functionality, usage, etc. For example, "Can I wear my Fitbit Alta in the shower?" is a typical factual question that can be directly answered based on product information.
- **User Information**: The user information such as previous purchases and browsing history are essential for the artificial personal shopper, as it would enable the inference of the context of the conversation and hence to provide a response that is personalized to individual users. For example, for the question "Is the Bose headphone compatible with my phone?", user information would allow the agent to infer that 'the Bose headphone' is 'a Bose QC35 headphone' by using search or browsing history, and 'my phone' is 'iPhone 6 (with iOS 10.3.2)' according to the purchase history.
- **Customer Generated Content**: Most online stores encourage customers to review and rate products, to submit product-related questions, and to answer other customer questions. In addition, customers can rate reviews and answers of other customers. This framework of customer generated content (CGC) complements the official information provided by the product provider and enables customers to take better shopping decisions by letting them learn from other customers' experience. Moreover, the CGC data can be used by the artificial shopper assistant for answering subjective questions asking for opinion or advice. For example, the question "Is iPad good for kids?" should be properly responded by extracting information from the iPad related reviews which discuss this particular topic. Typically, different opinions are expected for subjective topics, especially for the controversial ones, hence the agent's response should fairly cover the spectrum of the crowd opinions.

3 Research Challenges

In this section, we introduce research challenges (RCs) regarding how to handle the artificial personal shopper task. In addition, we discuss related work in IR and NLP that could be explored to tackle each of the research challenges.

RC1: How to process a voice utterance?

Advances in automatic speech recognition (ASR), especially with neural networks (e.g. Graves et al. 2013), enable an effective automatic transcription from voice to text utterances. Voice interaction opens many opportunity for search-based systems as users tend to provide more detailed questions as well as much more feedback for the search results (Guy, 2016). On the other hand, background noise, cross-talks, different accents, etc., cause many ASR errors. High-accuracy ASR is crucial for this task, as a small error could lead to an incomprehensible or misinterpreted transcribed utterance. Since the ASR technology is not perfect, a robust approach that provides a highly precise response for a noisy utterance is an important research challenge that has to be investigated.

RC2: How to identify appropriate response source(s) for a given utterance?

The optimal information sources for response generation should be identified according to the

Utterance	Potential Response
Is my S8 unlocked?	Yes, your Samsung Galaxy S8 is unlocked and can be used with any valid SIM card.
What is the best Kindle to buy?	Kindle PaperWhite is the top high-rated Kindle.
Tell me about Echo Dot.	Echo Dot is a smart speaker developed by Amazon ...
What is the best deal for Instant Pot?	Instant Pot DUO60 is currently 30% off.
Should I buy Galaxy S9 or iPhone 8?	Galaxy S9 has got higher ratings than iPhone 8 ...
Is iPad good for kids?	80% of our customers find that iPad 2017 is not good for kids, while 20% thought it is.
Does Anova Sous Vide make a lot of noise?	95% of our customers say that Anova Sous Vide is very silent in comparison with other products they have.
Is the Bose headphone compatible with my phone?	Bose QC35 headphone cannot be used with your iPhone 6.
I like this pair of Nike shoes.	Good choice. They are the top rating running shoes and match well with the running kit in your shopping cart.

Table 1: Examples of shopping related questions and potential responses from an artificial personal shopper.

utterance characteristics and type. For example, factoid questions should be better answered by the product source while advice questions should be answered by customer generated content. Identifying the proper response sources for a given utterance can be casted as a text categorization task that aims to label a natural language text with a category (or categories) in a pre-defined taxonomy of response sources.

While text categorization (or text classification) has been well-studied in the field of NLP and IR (e.g. Yang et al. 2016; Sebastiani 2002), it would be an interesting research challenge to develop a novel classifier and a set of features that could identify an appropriate response source effectively, while minimizing the risk of missing appropriate sources for the product domain. Another interesting challenge is how to optimally aggregate the results from different sources. For example, the quality of a question-answer pair can be evaluated according to the support it gets from related customer reviews (McAuley and Yang, 2016).

RC3: How to identify key phrases in a user utterance?

Previous work (e.g. Bendersky and Croft 2008) showed that only a few key terms or key phrases from a natural language query contribute significantly to the quality of the search results. For an artificial personal shopper, these key phrases in the user utterance are mainly the discussed products and their attributes which must be identified in order to support an effective conversation. Many existing techniques can be used to identify key phrases in a given text (e.g. Hulth 2003). However, existing key phrase extraction technologies were developed mainly for the general domain, such as websites or newswires, while limited work has been done in the product domain and in the noisy voice transcription domain. For example, an emphasis in the voice signal might be an indicator of a key-phrase. Hence, it is important to investigate into adapting existing approaches, and developing new domain specific approaches, to effectively extract key phrases from utterances for the artificial personal shopper task.

RC4: How to infer which product/entity the user refers to?

Another challenge of an artificial personal shopper is to infer which product or entity the user refers to. This is different from the traditional entity resolution task (e.g. Leidner et al. 2003) that mainly identify or match relevant entities in a pre-define ontology within the text. In the setting of an artificial personal shopper the entity resolution task is more complex since personalized information must be taken into consideration. For example, as already has been shown in Section 2, for the question "Is the Bose headphone compatible with my phone?", the system needs to infer that 'Bose headphone' is 'a Bose QC35 headphone' and 'my phone' is 'iPhone 6 (with iOS 10.3.2)', by using information from different sources, including browsing

history, purchase history, and the question itself. This would be an interesting area of research that needs to incorporate co-reference resolution and anaphora resolution for grounding personal shopper products/entities.

RC5: How to generate a natural language response?

Assuming that we could retrieve a piece of information that is relevant to the user utterance, the next major challenge is to generate a friendly conversational response that contains the relevant information as part of the continues dialog. Such a response should be comprehensive and complete while still concise and short. Several approaches could be investigated and extended for the task, including snippet generation (Turpin et al., 2007), text summarization (Spärck Jones, 2007) and natural language text generation (Wen et al., 2015). Nevertheless, a response from a snippet generation technique (Turpin et al., 2007) may be informational but non-conversational, while on the other hand, a response from a language generation technique (Wen et al., 2015) would be conversational but may not answer properly the user question. Therefore, an effective approach for generating informative and conversational responses is an interesting and open research challenge.

Another interesting aspect of this challenge is generating a multi-facet answer to a subjective question that represents the crowd's multi opinions respectfully and truthfully. In contrast to factoid questions, subjective questions can have many valid answers since there is no absolute ground truth. A multi-aspect answer shall cover the distribution of the crowd opinions over the answer aspect space. The final answer should represent the selected aspects with their accumulated sentiment as reflected in the CGC data.

RC6: How to evaluate an end-to-end personal shopper system?

The evaluation of conversational agents is a research area that has not attracted much attention by the research community. Goh et al. (2007) discusses the inappropriateness of existing IR measures for response quality evaluation, and calls for new standard measures and related considerations. Radziwill and Benton (2017) presents a literature review of quality issues with chatbots. Most evaluation approaches rely on having human evaluators provide their subjective views of the system's performance.

Another possible evaluation paradigm is based on n-gram similarities, such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), which are typically used for machine translation and text summarization tasks. Within this paradigm, a dialog system is evaluated based on the overlap between an n-gram set of its response and that of the ground truth (Papineni et al., 2002; Lin, 2004).

In contrast, question-answering evaluation has been studied extensively (Voorhees and Tice, 2000; Rajpurkar et al., 2016). However, these studies mainly focused on factoid questions. The TREC's LiveQA track (Agichtein et al., 2015) evaluated the ability of a QA system to answer complex Yahoo Answers questions in real time. Human editors judged the answer quality. In general, automatic answers quality was far from being satisfiable, compared to human answers.

Nevertheless, responses from an artificial personal shopper have to be conversational. Therefore, how to evaluate the responses based on the criteria of both the relevance toward the user's information needs and the replication of a human-like conversation would be an interesting research challenge.

4 Conclusions

We have introduced the personal shopper task for an intelligent virtual assistant, where the goal is to develop novel technologies to aid on-line voice shopping. In particular, we highlighted challenges of developing such a system and discussed how existing IR and NLP techniques could be adapted and extended to deal with challenges of the task. Achieving this task would pave a way for intelligent virtual assistants to perform more complex tasks in conversational search, and stimulate further research.

References

Eugene Agichtein, David Carmel, Dan Pelleg, Yuval Pinter, and Donna Harman. 2015. Overview of the trec 2015 liveqa track. In *TREC*.

Michael Bendersky and W. Bruce Croft. 2008. Discovering key concepts in verbose queries. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Infor-*

mation Retrieval, SIGIR '08, pages 491–498, New York, NY, USA. ACM.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*.

Mikhail Burtsev, Aleksandr Chuklin, Julia Kiseleva, and Alexey Borisov. 2017. Search-oriented conversational ai (scai). In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '17, pages 333–334, New York, NY, USA. ACM.

Ong Sing Goh, C. Ardil, W. Wong, and C.C. Fung. 2007. A black-box approach for response quality evaluation of conversational agent systems. *International Journal of Computational Intelligence*, pages 195–203.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE.

Ido Guy. 2016. Searching by talking: Analysis of voice queries on mobile web search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 35–44, New York, NY, USA. ACM.

Anette Hulth. 2003. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 216–223, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zongcheng Ji, Zhengdong Lu, and Hang Li. 2014. An information retrieval approach to short text conversation. *arXiv preprint arXiv:1408.6988*.

Jochen L Leidner, Gail Sinclair, and Bonnie Webber. 2003. Grounding spatial named entities for information extraction and question answering. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references-Volume 1*, pages 31–38.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Julian McAuley and Alex Yang. 2016. Addressing complex and subjective product-related queries with customer reviews. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 625–635, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Nicole M. Radziwill and Morgan C. Benton. 2017. Evaluating quality of chatbots and intelligent conversational agents. *CoRR*, abs/1704.04579.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natural language processing*, pages 583–593.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

Elizabeth Shriberg, Andreas Stolcke, Daniel Jurafsky, Noah Coccaro, Marie Meteer, Rebecca Bates, Paul Taylor, Klaus Ries, Rachel Martin, and Carol Van Ess-Dykema. 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and speech*, 41(3-4):443–492.

Karen Spärck Jones. 2007. Automatic summarising: The state of the art. *Inf. Process. Manage.*, 43(6):1449–1481.

Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2017. Misc: A data set of information-seeking conversations. In *SIGIR 1st International Workshop on Conversational Approaches to Information Retrieval (CAIR'17)*, volume 5.

Andrew Turpin, Yohannes Tsegay, David Hawking, and Hugh E Williams. 2007. Fast generation of result snippets in web search. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 127–134. ACM.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 200–207, New York, NY, USA. ACM.

Joseph Weizenbaum. 1966. Eliza a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1711–1721.

Jason D Williams and Steve Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.