Decision-support for the Masses by Enabling Conversations with Open Data

Biplav Srivastava
IBM

Abstract
Open data refers to data that is freely available for reuse. Although there has been rapid increase in availability of open data to public in the last decade, this has not translated into better decision-support tools for them. We propose intelligent conversation generators as a grand challenge that would automatically create data-driven conversation interfaces (CIs), also known as chatbots or dialog systems, from open data and deliver personalized analytical insights to users based on their contextual needs. Such generators will not only help bring Artificial Intelligence (AI)-based solutions for important societal problems to the masses but also advance AI by providing an integrative testbed for human-centric AI and filling gaps in the state-of-art towards this aim.

Introduction
As the world has shifted towards an increased digital economy and organizations have adopted the open data principles to make their data widely available for reuse \cite{herman2012data, Hughes2018}, there is an unprecedented opportunity to generate insights to improve the conditions of people around the world towards basic concerns of living like health, water, energy, traffic, community and environment \cite{Srivastava2015}. However, the current situation is that available interfaces of search and visualization for open data are targeted towards developers and are rudimentary when compared to what people want - relevant, prescriptive, data-driven insights for taking decisions. Building intelligent interfaces with open data needs AI and data specialization, is costly and consequently, slow to develop. The first glimpses can be seen in chatbot released by the state of Kansas \cite{bloomberg2017} in US while the private sector has tried to provide a natural question-answering interface \cite{sorkin2017} to US data. Research community has also started to build many prototypes in the area \cite{Ellis2018, Kephart2018}, and we seek to accelerate the phenomenon.

For people, natural interfaces like conversation have long been known to be effective. A conversation interfaces (CI), also known as chatbot or dialog system \cite{mctear2016}, is an automated agent, whether physical like robots or abstract like an avatar, that can not only interact with a person, but also take actions on their behalf and get things done. A simple taxonomy of interfaces we consider is shown in Table 1. One can talk to a chatbot or, if speech is not supported, type an utterance and get the systems response. They can be embedded along with other interaction modalities to give a rich user experience. The chatbot may converse without a goal in pleasantries and hence not need access to data sources, or be connected to a static data source like a company directory or a dynamic data source like weather forecast. The application scenarios become more compelling when the chatbot works in a dynamic environment, e.g., with sensor data, interacts with groups of people who come and go rather than only an individual at a time, and adapts its behavior to peculiarities of user(s).

There has been an upsurge in availability of chatbots for people on mobile phones and physical devices like Amazon Alexa and Google Home and numerous platforms have emerged to create them quickly for any domain \cite{Accenture2016}. Chatbots have been deployed in customer care in many industries where they are expected to save over $8 billion per annum by 2022 \cite{Juniper2017}. Chatbots can help users especially in unfamiliar domains when users do not know everything about the data, its metadata, kinds of analyses possible and the implications of using the insights.

However, the process to build them needs long development cycle and is costly. The main problem here is dialog management, i.e., creating dialog responses to user’s utterances by trying to understand user’s intent, determining the

| S.No. | Dimension | Variety |
|-------|-----------|---------|
| 1     | User      | 1, multiple |
| 2     | Modality  | only conversation, only speech, multi-modal (with point, map, …) |
| 3     | Data source | none, static, dynamic |
| 4     | Personalized | no, yes |
| 5     | Form      | virtual agent, physical device, robot |
| 6     | Purpose   | socialize, goal: information seeker, goal: action delegate |
| 7     | Domains   | general, health, water, traffic, … |

Table 1: Different Types of Conversation Interfaces

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
\footnote{https://toolbox.google.com/datasetsearch}
\footnote{Examples - http://data.gov, http://data.gov.in}
most suitable response, building queries to retrieve data and deciding the next course of action (responding, seeking more information or deferring). The chatbot so created also has to be tested for functional and non-functional characteristics, and social behavior.

We envisage a simpler and cost-efficient process where the user can point to a data-source like water quality and regulations, and gets a chatbot to ask whether they can safely drink a location’s water given others have done it on other days or locations. The user can then point to a disease data-source like water quality and get a chatbot to ask whether they can safely access to new domains and data sources based on a user’s needs. We envisage software programs, that we call chatbot generators, for generating conversation interfaces to deliver data-driven insights and also become personalized over time. The system would be able to quickly adapt to new domains and data sources based on a user’s needs, generate chatbots that are broadly useful, trustable by being able to transparently explain their decision process and awareness of fairness issues, and able to deploy chatbots widely in different forms (Table 1).

We now discuss open data, challenges in using them and some use-cases where insights from them can help people. Next, we review how conversation can help address the challenges and how the proposed chatbot generator can fill the gap for common usage patterns. We will use water as a case-study throughout to motivate how general people may benefit, where a prototypical multi-modal chatbot called Water Advisor (WA) was recently described (Ellis et al. 2018). We identify AI opportunities for learning, reasoning, representation and execution, along with human-centric design and ethics, to motivate more conversation applications.

Open Data and Its Challenges

Open data is an important trend in governance and computing over the last decade (Hughes and Rumsey 2018, W3C 2018). According to Open Data Catalogs, there are over 550 repositories, each with resources ranging from 10s to hundreds of thousands spanning most areas of human endeavor. The Open Data Barometer (W3C 2018) prepared a report surveying 1,725 datasets from 15 different sectors across 115 countries. In the first wave of the trend, which started around 2008, the focus was on acquiring data and making it available on scalable public platforms. In the second wave, the focus has been on driving consumption by providing richer semantics (Wright et al. 2012).

Many use-cases have been published demonstrating how insights from open data using AI methods can benefit people. For example, Srivastava (2015) gives a tutorial focusing on value of new AI insights in a domain, availability of relevant open data and context of people’s interaction with the (new) system. Consider water as an example. People make many daily decisions touching on water usage activities like for profession (e.g., fishing, irrigation, shipping), recreation (e.g., boating), wild life conservation (e.g., dolphins) or just regular living (e.g., drinking, bathing, washing). If accessible tools were available to public, they would be particularly useful to handle public health challenges such as the Flint water crisis (Pieper et al. 2017). The very few tools available today target water experts such as WaterLive mobile app for Australia, Bath app for UK, and GangaWatch for India (Sandha et al. 2017) and assume technical understanding of sciences.

Psychologists have long explored the sense-making process of analysts when looking at data through cognitive task analysis of their activities (Pirolli and Card 2004). They found that analysts try to explore schema and query data, look at data, build hypothesis and look for evidence. The current interfaces to access data are intended for developers and data analysts. They consist of search on repository sites (like data.gov) or via search engines and application programming interfaces (APIs). Further, for data analysis, analysts want to understand the context of data available, the standards and issues prevailing in a domain of interest and a forum to discuss inter-disciplinary challenges. In fact, in John et al. (2017), the authors have created a chatbot to help with data analysis steps.

But our focus is on how general public may benefit from insights generated from open data without long development cycles and in context of their use. The proposed chatbot generator will enable a new, complimentary, conversational, interface to open data that a user can interact, possibly as part of a multi-modal user experience. Thus, a user will be able to talk to an automated chatbot to get insight she wants while optionally also seeing additional, relevant, visualizations and documents the agent may be able to retrieve.

Conversation Interfaces

There is a long history of Conversational interfaces (CIs) going back to the 1960s when they first appeared to answer questions or do casual conversation (McTear et al. 2016). A conversation, or dialog, is made up of a series of turns, where each turn is a series of utterances by one or more participants playing one or more roles. A common type of chatbot deals with a single user at a time and conducts informal conversation, answers the user’s questions or provides recommendations in a given domain. It needs to handle uncertainties related to human behavior and natural language, while conducting dialogs to achieve system goals.

Building Data-Consuming Chatbots

The core problem in building chatbots is that of dialog management, i.e., creating dialog responses to user’s utterances. The system architecture of a typical data-consuming dialog manager (DM) is shown in Figure 1. Given the user’s utterance, it is analyzed to detect her intent and a policy for handling the next turn is selected. This policy may call for querying a database, and the result is returned which is used by response generator to create a response using templates. The system

---

Footnotes:

1. https://datacatalogs.org/. Accessed 12-Sep-2018.
2. http://www.water.nsw.gov.au/realtime-data
3. https://environment.data.gov.uk/bwq/profiles/
4. https://toolbox.google.com/datasetsearch
5. See examples at http://data.gov, http://data.gov.in.
can dynamically create one or more queries which involves selecting tables and attributes, filtering values and testing for conditions, and assume defaults for missing values. It may also decide not to answer a request if it is unsure of a query’s result correctness.

Note that the DM may use one or more domain-specific data bases (sources) as well as one or more domain-independent sources like language models and word embeddings. Common chatbots use static domain-dependent databases like product catalogs or user manuals. The application scenarios become more compelling when the chatbot works in a dynamic environment, e.g., with sensor data, and interacts with groups of people, who come and go, rather than only an individual at a time. In such situations, the agent has to execute actions to monitor the environment, model different users engaged in conversation over time and track their intents, learn patterns and represent them, reason about best course of action given goals and system state, and execute conversation or other multi-modal actions.

There are many approaches to tackle DM in literature including finite-space, frame-based, inference-based and statistical learning-based (Crook 2018; Clark et al. 2010; Inouye 2004; Young et al. 2015) of which finite-space and frame-based are most popular with mainstream developers. Task-oriented DMs have traditionally been built using rules. Further, a DM contains several independent modules which are optimized separately, relying on huge amount of human engineering. The recent trend is to train DM from end-to-end, allowing error signal from the end output of DM to be back-propagated to raw input, so that the whole DM can be jointly optimized (Bordes et al. 2017). A recent paper reviews the state-of-art and looks at requirements and design options to make them customizable for end users as their own personal bot (Daniel et al. 2018).

There are also unique considerations when exploring data with dialog:

• Dynamic source selection: the data in a domain may consist of multiple tables, attributes (columns) and rows. The user utterance could lead to discovering them and they then become part of the context of query.
• Query patterns: there are often common patterns and natural order of user queries. Users may adopt them to explore data and can be used as shared context.
• Query cost: the order of query execution could be important for cost reasons.
• Query mapping: mappings from natural language to data model may have to be learned and adapted based on different source models.
• Conversation length: As conversations become long, there could be increased risk of the user dropping off leading to diminishing returns.

**Usability Opportunities and Issues with Chatbots**

An obvious question to ask when considering chatbots is when they are most suitable. The effectiveness of conversation versus other modalities has long been studied (Frohlich 1993). Some scenarios where conversation is quite suitable include when users are unfamiliar with the domain, expect non-human actors due to unique form of agent embodiment (e.g., robots) or prefer them (e.g., due to sensitivity of the subject matter), and where content changes often and users seek guidance (Srivastava 2017).

However, such systems can also be fraught with ethical risks. An extreme and anecdotal example was the Tay (Neff and Nagy 2017) system in 2016 that was designed to engage with people on open topics over Twitter and learn from feedback, but ended up getting manipulated by users to exhibit unacceptable behavior via its extreme responses. The authors in (Henderson et al. 2018) systematically identify a number of potential ethical issues in dialogue systems built using learning methods: showing implicit biases from data, being prone to adversarial examples, vulnerable to violate privacy, need to maintain safety of people, and concerns about explainability of response and reproducibility of results. To handle these concerns, more research is needed. One idea here is to augment open data repositories that usually consist of standard information like data and size, usage license (context), how data was obtained (provenance), semantics of missing values and units, and a responsible person. Researchers have proposed to further describe protected variables and fairness considerations as datasheets (Gebru et al. 2018). The metadata can be used along with recent techniques and tools to address issues of fairness and transparency with chatbots (Henderson et al. 2018) and build trust with users.

**On Use-Cases With Open Data And Dialogs**

Over the years, a number of common analysis patterns have emerged across domains which have been shown to be useful to people around the world (Srivastava 2015). One pattern is Return of Investment. Here, the monetary investment made into a domain is compared against suitable metrics of outcome and improvement is sought. As example, public funds invested into water work may be analyzed to see reduction in cases of heavy metal contamination, health care may be analyzed to see reduction in number of patients and deaths, or funds invested for tourist promotion be compared with increase in economic activity in a city. Another pattern is Comparison of Results. Here, if an improvement is

---

8 AI Fairness Toolkit - https://github.com/IBM/AIF360
used to discovering issues in water quality together with regulation data. More generally, learning can be used for alternative DM approaches like end-to-end policy learning from data (Bordes et al. 2017).

**Representation** is needed to model location, time and user. For water, it encodes regulation and safe limits and mapping of usage purpose to quality parameters.

**Reasoning** is crucial to keep conversation focused based on system usability goals and user needs. One can model cognitive costs to user based on alternative system response choices and seek to optimize short-term and long-term behavior. For water, reasoning is used to short-list regulations based on water activity and region of interest, generate advice and track explanations.

**Execution** is autonomous as the agent can choose to act by (a) asking clarifying questions about user intent or locations, (b) asking user’s preference about advice, (c) seeking most reliable data source (water) for region and time interval of interest from available external data sources, and corresponding subset of compatible other sources (regulations) (d) invoking reasoning to generate an advice (for water usage using filtered water data and regulations), (e) visualizing its output and advice, and (f) using one or more suitable modalities available at any turn of user interaction, i.e., chat, maps and document views.

**Human Usability Factors** have to be explicitly modeled and supported. The user-interface controller module automatically keeps the different modalities synchronized and is aware of missing data or assumptions it is making, so that they are used while communicating output advice in generated natural language. Further extensions can be to measure and track complexity of interaction (Liao et al. 2017) and use sensed signals to pro-actively improve user experience and combine close-ended and open-ended questioning strategies for efficient interaction (Zhang et al. 2018).

**Ethical Issues** can emerge whenever a piece of technology is used among people at large. We discussed handling them in a domain independent manner earlier. There may also use-cases needing domain-specific considerations due to which scope of chatbot generator has to be selectively expanded.

### Conclusion

In this paper, we proposed the challenge of **intelligent conversation interface generator** that, given a set of open data sources of interest, would generate a chatbot that can interact autonomously with a common person (non-developer) and provide insights about selected data. Such a technology would bring Artificial Intelligence (AI)-based solutions for important societal problems to masses along common patterns of usage while technical AI experts focus on specialized cases. It will also serve as an integrative testbed for human-centric AI and advance the sub-areas.

### References

[Accenture 2016] Accenture. Chatbots in customer service. In At: https://accntu.re/2z9s5fH, 2016.
[Bloomberg 2017] Bloomberg. Chatbot makes open data user-friendly. In https://medium.com/@BloombergCities/chatbot-makes-open-data-user-friendly-c1888c863db4, 2017.

[Bordes et al. 2017] Antoine Bordes, Y-Lan Boureau, and Jason Weston. Learning end-to-end goal-oriented dialog. In Proc. ICLR, 2017.

[Clark et al. 2010] Alexander Clark, Chris Fox, and Shalom Lappin. Handbook of computational linguistics and natural language processing. In Wiley, ISBN: 978-1-405-15581-6, 2010.

[Crook 2018] Paul Crook. Statistical machine learning for dialog management: its history and future promise. In AAAI DEEP-DIAL 2018 Workshop, at https://www.dropbox.com/home/AAAI2018-DEEPDIALWorkshop/Presentations-Shareable?preview=Invited1-PaulCrook-AAAI_DeeperDialog_Feb2018.pdf, 2018.

[Daniel et al. 2018] Florian Daniel, Maristella Matera, Vittorio Zaccaria, and Alessandro Dell’Orto. Toward truly personal chatbots: On the development of custom conversational assistants. In Proceedings of the 1st International Workshop on Software Engineering for Cognitive Services, SE4COG ’18, pages 31–36, New York, NY, USA, 2018. ACM.

[Ellis et al. 2018] Jason Ellis, Biplav Srivastava, Rachel Bellamy, and Andy Aaron. Water advisor - a data-driven, multimodal, contextual assistant to help with water usage decisions. In Proc. 32nd AAAI Conference on Artificial Intelligence (AAAI-18), New Orleans, Louisiana, USA., 2018.

[Frohlich 1993] David Frohlich. The history and future of direct manipulation. In HP Tech Report, http://www.hpl.hp.com/techreports/93/HPL-93-47.pdf, 1993.

[Gebru et al. 2018] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. CoRR, abs/1803.09010, 2018.

[Henderson et al. 2018] Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. Ethical challenges in data-driven dialogue systems. In Proc. of AAAI/ACM Conference on AI Ethics and Society (AIES-18), New Orleans, Louisiana, USA., 2018.

[Hereman and Beeman 2012] I. Herman and H. Beeman. Open data in practice. In W3C Tutorial Track, 21st International World Wide Web Conference (WWW2012), Lyon, France, 2012.

[Hughes and Rumsey 2018] Adam Hughes and Matt Rumsey. The state of the union for open data. In Data Foundation Report, At http://www.datafoundation.org/the-state-of-the-union-of-open-data-ed-2/, 2018.

[Inouye 2004] R. Bryce Inouye. Minimizing the length of non-mixed initiative dialogs. In Daniel Midgley Leonoor van der Beek, Dmitriy Genzel, editor, ACL 2004: Student Research Workshop, pages 7–12, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[John et al. 2017] Rogers Jeffrey Leo John, Navneet Potti, and Jignesh M. Patel. Ava: From data to insights through conversations. In CIDR, 2017.

[Juniper 2017] Juniper. Chatbots: Retail, ecommerce, banking & healthcare 2017-2022. In At: http://bit.ly/2sJHejY, 2017.

[Kephart et al. 2018] Jeffrey Kephart, Victor Dibia, Jason Ellis, Biplav Srivastava, Kartik Talamadupula, and Mishal Dholakia. Cognitive assistant for visualizing and analyzing exoplanets. In Proc. 32nd AAAI Conference on Artificial Intelligence (AAAI-18), New Orleans, Louisiana, USA., 2018.

[Liao et al. 2017] Q. Liao, B. Srivastava, and P. Kapani pathi. A Measure for Dialog Complexity and its Application in Streamlining Service Operations. ArXiv e-prints, August 2017.

[McTear et al. 2016] M. McTear, Z. Callejas, and D. Griol. Conversational interfaces: Past and present. In The Conversational Interface. Springer, DOI: https://doi.org/10.1007/978-3-319-32967-3_4, 2016.

[Neff and Nagy 2017] G. Neff and P. Nagy. Automation, algorithms, and politics—talking to bots: Symbiotic agency and the case of tay. In International Journal Of Communication, 10, 17. Retrieved from http://ijoc.org/index.php, 2017.

[Pieper et al. 2017] Kelsey J. Pieper, Min Tang, and Marc A. Edwards. Flint water crisis caused by interrupted corrosion control: Investigating “ground zero” home. Environmental Science & Technology, 51(4):2007–2014, 2017.

[Pirolli and Card 2004] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In Proceedings of International Conference on Intelligence Analysis, 2004.

[Sandha et al. 2017] Sandeep Singh Sandha, Biplav Srivastava, and Sukanya Randhawa. The gangwatch mobile app to enable usage of water data in every day decisions integrating historical and real-time sensing data. CoRR, abs/1701.08212, 2017.

[Sorkin 2017] Andrew Ross Sorkin. Steve ballmer serves up a fascinating data trove. In https://www.nytimes.com/2017/04/17/business/dealbook/steve-ballmer-serves-up-a-fascinating-data-trove.html, 2017.

[Srivastava 2015] Biplav Srivastava. Ai for smart city innovations with open data. In Tutorial at International Joint Conference on Artificial Intelligence (IJCAI), at Buenos Aires, Argentina. Details: https://sites.google.com/site/aismartcitytutorial/, 2015.

[Srivastava 2017] B. Srivastava. Designing talking alecks that look smart and make users happy. In https://www.linkedin.com/pulse/designing-talking-alecks-that-look-smart-make-users-happy-srivastava-, 2017.

[W3C 2018] W3C. Open data barometer, 4th ed. In At https://opendatabarometer.org/4thedition/report/, 2018.

[Wright et al. 2012] Glover Wright, Praneeth Prakash, Sunil Abraham, and Nishant Shah. Open data in india. In CIS Report at http://cis-india.org/publications/ogd-report, 2012.

[Young et al. 2013] Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. Pompdb-based statistical
spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179, 2013.

[Zhang et al., 2018] Yunfeng Zhang, Vera Liao, and Biplav Srivastava. Towards an optimal dialog strategy for information retrieval using both open-ended and close-ended questions. In *Proc. Intelligent User Interfaces (IUI 2018, Tokyo, Japan, March)*, 2018.