On the Granularity of Dialog Strategies: Insights from Large-scale Analyses of Two Commercial Travel Information Spoken Dial....

Conference Paper · March 2015

CITATIONS
0

READS
138

6 authors, including:

Zhuoran Wang
Tricorn (Beijing) Technology Co., Ltd. (trading as trio.ai)
27 PUBLICATIONS **195** CITATIONS

SEE PROFILE

On the Granularity of Dialog Strategies: Insights from Large-scale Analyses of Two Commercial Travel Information Spoken Dialog Systems

Zengtao Jiao[1], **Zhuoran Wang**[2,3], **Guanchun Wang**[1], **Hao Tian**[1], **Hua Wu**[1], **Haifeng Wang**[1]
[1]Baidu Inc., No. 10, Shangdi 10th Street, Haidian District, Beijing, P. R. China
[2]Heriot-Watt University, Edinburgh, UK [3]Toshiba Research Europe Ltd., Cambridge, UK

Abstract

This paper empirically analyzes large amounts of real user dialogs collected from two deployments of spoken dialog systems (SDS) in a travel information mobile app widely used in China. The data studied here consist of 288K dialogs from a flight booking SDS and 2.42M dialogs from its extended version that handles both flight and hotel reservations. Based on such data, we investigate user behaviors and the underlying intentions behind their behaviors. We also examine how such user intentions/behaviors vary with different location-related and time-of-day-related factors. The findings in this research can be used to design more granular dialog strategies and allow richer dialog state featurizations in order to better address real user demands for SDS in similar domains.

Introduction

With growing demands for natural human-machine interfaces, spoken dialog systems (SDS) have been increasingly deployed in various commercial applications over the last decade. Recent advances in SDS have shown that statistical approaches such as partially observable Markov decision processes (POMDPs) in conjunction with reinforcement learning can result in significantly more robust and natural dialogs (Young et al. 2013). The dialog policies in such statistical SDS are usually learnt based on either user simulations (Schatzmann et al. 2007; Young et al. 2010) or crowdworkers (Gasic et al. 2013). However, the intentions and behaviors of real users may vary with many environmental, regional or sociological factors, which poses the following high-level questions: (1) To what granularity should dialog policies be designed in order to capture the diversity of real user intentions in real world scenarios? (2) What external factors in addition to observed dialogs themselves could help an SDS to discriminate such diverse user intentions?

To address the above questions, we empirically analyze two large collections of real user dialogs obtained respectively from two travel information SDS publicly deployed in China. The purpose of the analyses here is to better understand hidden user intentions embedded in (mostly implicit) user behavior patterns. Several interesting phenomena are identified, based on which we discuss possible ways of using consequential findings to improve the granularity design of dialog modeling for SDS in similar domains.

Comparing to previous empirical SDS studies based on public deployments (Williams 2011; 2012), this work is focused on user behaviors instead of model-related system characteristics. To the best of our knowledge, this is the first paper on human subject research for SDS that investigates datasets as large as millions of dialogs.

Descriptions of the Systems and Data

The two SDS used for data collection in this paper correspond to two deployments of a travel assistant mobile app, developed by Baidu and widely used in China. The first system (SDS1) is designed for flight booking only, while the second system (SDS2) extends the application domains to flight booking and hotel reservation. Both SDS are rule-based, following the RavenClaw architecture (Bohus and Rudnicky 2009), and are in Chinese language. One-best ASR hypotheses are used in both systems when making decisions, where the ASR word accuracy is 90.0%. The SDS can provide external links for actual order placements, but cannot access further transaction information.

The fundamental dialog management mechanism in our SDS can be interpreted as a slot-filling procedure, where required slots and optional slots are defined for each task separately. In every dialog, all the required slots must be filled before the system can execute a database search and display the query results to the user. For flight booking, there are 3 required slots, namely departure city, destination city and departure date, and 14 optional slots such as time of day, seat grade, airline, etc. For hotel reservation, the 3 required slots are destination city, check-in and check-out dates, with 9 optional slots including landmark, star-rating, etc.

The data under study in this paper consist of 288K dialogs with 610K turns collected from SDS1 during January to May 2014 (denoted as DS1) and 2.42M dialogs with 5.72M turns collected from SDS2 during January to June 2014 (denoted as DS2). In DS2, 90% of the dialogs are for single tasks, with 60% for flight booking and 30% for hotel reservation. There are 2% of the dialogs in DS2 that contain compound flight and hotel booking goals, while no clear user goals are identifiable in the remaining 8% of the dialogs in DS2. The

estimated task completion (ETC) rate[1] for SDS1 and SDS2 are 77.0% and 78.5%, respectively.

What makes dialogs fail?

We start from some simple statistics to examine the missing (required) slot information that results in dialog failures in our SDS, of which the results are shown in Table 1[2].

	Flight		
	Departure	Destination	Date
DS1	0.087	0.394	0.578
DS2	0.115	0.214	0.750
	Hotel		
	Destination	Check-in	Check-out
DS2	0.214	0.749	0.722

Table 1: The proportions of missing values for required slots in the incomplete dialogs in DS1 and DS2.

Firstly, it can be found that the lack of date information results in most of the dialog failures in both systems and for both tasks. This is intuitively understandable, as when planning a trip, a user should already have his/her departure and destination in mind, but may not always have a clear idea of the exact travel date. In many cases, users may just want to browse through possible schedules and compare the prices of flight tickets or hotel rooms. Nevertheless, defining those required slots is a necessary mechanism to refine database queries for the ease of information access. On the other hand, in many situations, e.g. when user goal is clear and fixed or when time cost is of concern, users may prefer more brief dialogs than continuous browsing. Better dialog strategies should be able to identify different user intentions and to address them in corresponding manners.

It poses a further question here that how often users tend to explicitly state their intentions. Taking the intention of browsing travel plans as an example, by identifying relevant user acts using template-based approaches, we find that, in both DS1 and DS2, there are only around 0.11% of the dialogs that contain explicit user statements of "date unplanned" or "to browse". Similarly and probably even worse, one can hardly expect a user to express his/her demand of brief interactions by explicitly urging the system. Such implicit user behaviors suggest that to obtain more granular dialog strategies, extra knowledge or factors should be considered to enable the system to infer hidden user intentions.

[1]Throughout this paper, *estimated task completion rate* stands for the percentage of dialogs where all required slots are filled. The ETC rate reflects the ability and willingness of the users to complete their travel booking tasks with the SDS. Since this paper aims to analyze macro-level patterns of user behaviors, it is impractical to compute true user goal satisfaction rates by manually labeling the data.

[2]Note that, as our SDS will initialize the departure place according to user's GPS location if such information is available, most of those incomplete dialogs can still have their departure slots filled by default. Therefore, in Table 1, the proportions of the incomplete dialogs caused by missing departure values are much lower than those caused by missing other slot values.

What factors may imply user intentions?

Intuitively, one could imagine that a user's destination will to some extent reflect the purpose of the trip. In addition, the way that a user interacts with an SDS may also vary during different time periods of a day (e.g. busy hours and leisure hours). Therefore, we examine the correlations of user intentions with location and time-of-day related factors as follows.

Location-specific characteristics

Starting from the flight booking problem, we analyze the correlations between ETC rates and user's departure and destination cities in Table 2, where the most interesting findings are highlighted.

There is a common phenomenon in both datasets that, some most popular tourist cities[3] in China, such as Lhasa, Sanya, Lijiang, etc., demonstrate exactly opposite effects to the ETC rates when being the departure or the destination locations. To explain this, one can imagine that a user using the flight booking system at a tourist place would tend to have a clear goal in mind (e.g. searching for a flight back home), whilst in many cases the users searching for flights to tourist places may just want to browse flights and to compare prices without any specific fixed plan, especially when it comes to the travel date.

As the true browsing user intentions are unobservable, we reflect the above assumption from the following perspectives. Firstly, for each destination city, we compute the percentage of the incomplete dialogs where date values are missing (named *date missing rate*) and the percentage of the completed dialogs where users continue the dialogs (e.g. to search for alternative options) after the results for their first queries being displayed (named *perceived browsing rate*), of which the results are illustrated in Table 3. Note here, we only show the statistics based on DS2, as the examples of interest in DS1 are too sparse. It can be found that the lack of date information results in more frequent dialog failures for the users traveling to those tourist cities than for users to most of the other destinations, while the users going to those tourist destinations tend to browse more often for alternative options after their initial goals were satisfied. In addition, when those tourist cities are the perceived destinations, the percentage of dialogs where users explicitly state "date unplanned" or "to browse" increases to around 0.15%, which is slightly higher than the overall level (0.11%), though the absolute occurrence frequency of such user acts is still low.

All the above findings suggest that location-specific characteristics of user's destination, such as whether it is a popular tourism place, can reflect user's intention. Taking such information into account when designing dialog strategies could improve the performance and/or usability of the SDS. Such strategies could also generalize to destination cities

[3]All the highlighted cities are famous for the landscapes surrounding them, where there is no other notable industry or business except tourism.

DS1 Departure			DS1 Destination			DS2 Departure			DS2 Destination		
Rank	City	ETC	Rank	City	ETC	Rank	City	ETC	Rank	City	ETC
1	Hong Kong	1.000	1	Taipei	1.000	1	**Xidai***	0.995	1	Chongqing	0.997
2	**Lhasa**	0.999	2	Jinjiang	1.000	2	Hong Kong	0.993	2	Taipei	0.996
3	Jieyang	0.982	3	Chongqing	1.000	3	**Lhasa**	0.991	3	Qingdao	0.996
4	**Guilin**	0.955	4	Xi'an	1.000	4	**Kunming**	0.970	4	Dalian	0.996
5	**Kunming**	0.954	5	Shanghai	1.000	5	**Haikou**	0.967	5	Xi'an	0.996
6	Urumqi	0.950	6	Beijing	1.000	6	**Lijiang**	0.965	6	Shenyang	0.996
7	**Haikou**	0.950	7	Lanzhou	1.000	7	Xieyang	0.965	7	Harbin	0.996
8	Yinchuan	0.943	8	Hong Kong	1.000	8	Yinchuan	0.959	8	Nanjing	0.995
9	**Lijiang**	0.938	9	Shenyang	1.000	9	**Sanya**	0.959	9	Lanzhou	0.995
...	...	...	...	...	...	10	**Guilin**	0.959	10	Tianjin	0.995
13	**Sanya**	0.930				...	...	...	...	...	...
...	...	...	41	**Kunming**	0.994				43	**Haikou**	0.989
			43	**Haikou**	0.994				...	...	...
			...	...	...	103	Dongguan	0.822	53	**Sanya**	0.982
50	Taizhou	0.669	46	**Xidai***	0.991	104	Cixi	0.817	54	**Lijiang**	0.980
51	Yiwu	0.663	47	**Sanya**	0.984	105	Nan'an	0.810	55	**Zhangjiajie**	0.975
52	Suzhou	0.550	48	**Lijiang**	0.968	106	Fuqing	0.808	56	**Dali**	0.971
53	Dongguan	0.342	49	**Lhasa**	0.955	107	Foshan	0.804	57	**Lhasa**	0.965
Total: 53			Total: 49			Total: 107			Total: 57		

Table 2: ETC rates of flight departure and destination cities in DS1 and DS2. Cities with less than 500 and 1000 occurrences in DS1 and DS2 are filtered out, respectively. (* Xidai is the abbreviation for Xishuangbanna, or also known as Sibsongbanna.)

Date Missing			Perceived Browsing		
Rank	City	Rate	Rank	City	Rate
1	Huangshan	0.958	2	Dali	0.186
3	Dali	0.935	6	Zhangjiajie	0.173
7	Lijiang	0.902	7	Lijiang	0.166
8	Lhasa	0.898	8	Sanya	0.153
10	Zhangjiajie	0.880	10	Lhasa	0.150
Overall		0.755	Overall		0.126

Table 3: Date missing rates and perceived browsing rates for some highlighted tourist destinations in DS2 (in comparison to the respective overall rates for all applicable dialogs).

that have other specific types of characteristics. For example, users going to industry-intensive cities would usually have a clear travel plan in mind, and therefore might prefer more efficient travel booking procedures than browsing. However, many big cities may have simultaneous business and tourism functions. In such cases, it might be too assertive to estimate user intentions predictively. Nevertheless, instead of "guessing" the purpose of a user's trip, asking straightforwardly (e.g. "Okay, flight to Lhasa. Is it for business or leisure?") could be more preferable. But a more elegant SDS should be able to decide whether or when to ask user's travel purpose according to where the user is going to. Personalized statistics will also be helpful for such decisions.

Remark 1 *One might argue that this is a trivial problem, as state-of-the-art statistical SDS can learn such decisions using reinforcement learning by interacting directly with human users (Gasic et al. 2013). However, due to the commonly used summary space methods (e.g. (Young et al. 2010; Thomson and Young 2010)), the discriminations of individual slot values are usually eliminated when learning dialog policies. In addition, the extent to which crowdsourcing testers will behave the same as real users is also an open question. Therefore, improving the granularity of dialog state representations and dialog strategy designs in travel information SDS by considering location-specific characteristics is one of the main suggestions of this research.*

Remark 2 *Although a proper rule-based implementation (such as ours) can already result in rather high overall ETC rates, improving the granularity of dialog strategies will still be an issue of significant importance, since ETC does not directly relate to user satisfaction. Moreover, recommending options in a more suitable manner to the users who do not have clear travel plans is an indispensable function for a commercial travel information application.*

However, the above phenomena are domain-dependent. The statistics in hotel reservation dialogs indicate a different pattern of user behaviors, where users targeting tourist cities demonstrate higher ETC rates but less browsing intentions. This could be partially due to the fact that the prices of hotels usually do not change as much as the prices of flight tickets do on different dates.

Time-of-day-related factors

As mentioned above, different time periods during a day could also be the factors that affect user's behaviors when interacting with an SDS. We plot the task-dependent ETC rates

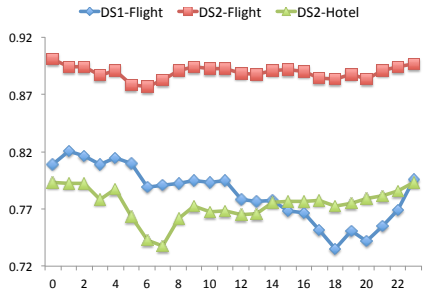

Figure 1: ETC rates with respect to daily time periods (hours).

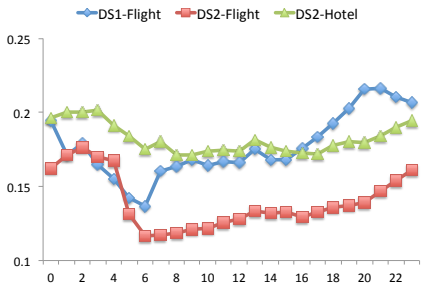

Figure 2: Perceived browsing rate with respect to daily time periods (hours).

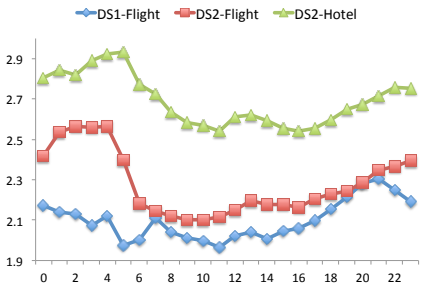

Figure 3: Average dialog lengths with respect to daily time periods (hours).

in DS1 and DS2 with respect to daily time periods (hours) in Figure 1.

For flight booking and hotel reservation dialogs in DS2, it can be found that the lowest ETC rates occurs in early morning around 6∼7am, while for DS1 (flight booking dialogs only) there is also a significant decrease in ETC rate around 6am, though the relative trend appears less obvious than that in DS2. More interestingly, highest ETC rates for all the tasks and systems are observed around midnight. Possible explanations to these phenomena could be as follows. Firstly, it can be understood that people using the system in "abnormal" time periods (such as the midnight) may have strong requirements and motivations to have their journeys/hotels booked. On the contrary, in the early morning, users may prefer more efficient or brief dialogs, and therefore tend to be less patient with the systems.

To prove our above assumptions, we carry out a time-dependent calculation of the perceived browsing rates and the average lengths of the dialogs as shown in Figure 2 and 3, respectively. Both statistics indicate that shorter dialogs and less browsing operations occur in the morning, while the opposite user behaviors are observed at late night.

Remark 3 *To the best of our knowledge, time-of-day-related factors have never been discussed in previous work on SDS. But the findings here suggest that including such information in dialog strategy design and/or for dialog policy learning could be a possible direction to improve the usability of SDS.*

In addition, one can find in Figure 1 that there are notable drops in the ETC rates for flight booking dialogs in both DS1 and DS2 during 6pm to 8pm. Users with uncertain goals trying to browse for travel plan options could be the main reason, as obvious peaks can be observed for DS1 in both Figure 2 and 3. However, there is no clear evidence to explain the same phenomenon in DS2. One possibility could be that the proportion of mistakenly triggered dialogs might increase during those "entertaining" hours.

Conclusion

Based on large-scale analyses of real user behaviors observed from two commercial travel information spoken dialog systems, this paper proposes several possible improvements to the granularity design of dialog strategy and dialog modeling for SDS in similar domains. The main contributions of this work are the investigations on the correlations between user behaviors and location-specific characteristics, as well as time-of-day-related factors. The feasibilities of the proposed improvements are also discussed based on empirical evidences. Practical implementations of SDS following such instructions to better address real user demands in real world scenarios will be the focus of our future research.

Acknowledgements

The research in this paper is supported by China's 973 Program (No. 2014CB340505). ZW is supported in part by a SICSA PECE grant.

References

Bohus, D., and Rudnicky, A. I. 2009. The RavenClaw dialog management framework: Architecture and systems. *Comp. Speech Lang.* 23(3):332–361.

Gasic, M.; Breslin, C.; Henderson, M.; Kim, D.; Szummer, M.; Thomson, B.; Tsiakoulis, P.; and Young, S. 2013. Online policy optimisation of Bayesian spoken dialogue systems via human interaction. In *ICASSP*.

Schatzmann, J.; Thomson, B.; Weilhammer, K.; Ye, H.; and Young, S. 2007. Agenda-based user simulation for bootstrapping a POMDP dialogue system. In *NAACL-HLT*.

Thomson, B., and Young, S. 2010. Bayesian update of dialogue state: A POMDP framework for spoken dialogue systems. *Comp. Speech Lang.* 24(4):562–588.

Williams, J. D. 2011. An empirical evaluation of a statistical dialog system in public use. In *SIGDIAL*.

Williams, J. D. 2012. Challenges and opportunities for state tracking in statistical spoken dialog systems: Results from two public deployments. *IEEE J. Selected Topics Sig. Proc.* 6(8):959–970.

Young, S.; Gasic, M.; Keizer, S.; Mairesse, F.; Schatzmann, J.; Thomson, B.; and Yu, K. 2010. The Hidden Information State model: a practical framework for POMDP-based spoken dialogue management. *Comp. Speech Lang.* 24(2):150–174.

Young, S.; Gasic, M.; Thomson, B.; and Williams, J. 2013. POMDP-based statistical spoken dialogue systems: a review. *Proc. IEEE* PP(99):1–20.