# Large-scale Analysis of the Flight Booking Spoken Dialog System in a Commercial Travel Information Mobile App

**Zengtao Jiao**
Baidu Inc.
Beijing, China

**Zhuoran Wang**
Heriot-Watt University
Edinburgh, UK

**Guanchun Wang**
Baidu Inc.
Beijing, China

**Hao Tian**
Baidu Inc.
Beijing, China

**Hua Wu**
Baidu Inc.
Beijing, China

**Haifeng Wang**
Baidu Inc.
Beijing, China

## Abstract

In this paper, we analyze around three hundred thousand real user dialogs collected from a publicly deployed flight booking spoken dialog system (SDS), to investigate the correlations between the task completion rate and user locations and daily time periods, as well as the correspondences between user responses and system requests. The findings can serve as guidelines to design more granular strategies for SDS in this domain.

## 1 Introduction

Due to the recent advances of mobile technology and the prevalence of smart devices in the latest decade, commercialized speech interfaces, particularly spoken dialog systems (SDS), are gaining increasing popularity. Successful examples include Apple's Siri, Google Now and Microsoft Cortana, to name just a few. The broad deployment of such applications enables more advanced analyses of SDS based on a vast amount of data generated by real users in real-world scenarios. Previous studies of this kind can be found in (Williams, 2011; Williams, 2012).

This paper studies several interesting phenomena observed in a large-scale data set of real user dialogs collected from a flight booking SDS developed by Baidu and integrated in a travel information mobile app widely used in China. The SDS here is a rule-based system following the RavenClaw architecture (Bohus and Rudnicky, 2009), where the dialog manager only takes top SLU hypotheses into account when making decisions.

## 2 Data Analysis

The data analyzed in this work consist of around 300K dialog sessions and more than 600K turns collected from our SDS during the first half of 2014. Basic statistics show that the task completion rate[1] for these dialogs is 77%. Based on such data, two factors that may affect the task completion rate are investigated in detail, including user's departure/destination locations and time periods of a day when the dialogs occur. In addition, we also investigate the correspondences between user responses and systems requests, which reflects user habits and the properness of each system action.

**Departures and Destinations** We cluster user's departure and destination cities according to the provinces they belong to, and plot the province-wise departure and destination task completion rates in Figure 1. Firstly and very interestingly, it can be found that the three most popular tourist provinces, Hainan, Yunnan and Tibet, demonstrate exactly opposite effects to the task completion rate when they occur as the departure and the destination locations. To explain this phenomenon, 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 a tourist place may just want to browse flights and compare prices without any specific plan in mind, especially the travel date (which results in 57.8% of the task failures according to our statistics). It suggests that a better dialog policy should consider user intentions adaptively when knowing the above prior knowledge, rather than treat all the destinations uniformly. Secondly, for some provinces, such as Hebei and Anhui, the task completion rate is relatively low, regardless of them being the departure or the destination locations. This may be because the ASR is less ro-

[1] There are 3 required slots (departure, destination and flight date) in the SDS, which must be filled before the system can execute a database search. The task completion rate defined here stands for the percentage of dialogs where all the three required slots are filled. There are 14 optional slots not considered when computing the task completion rate here.

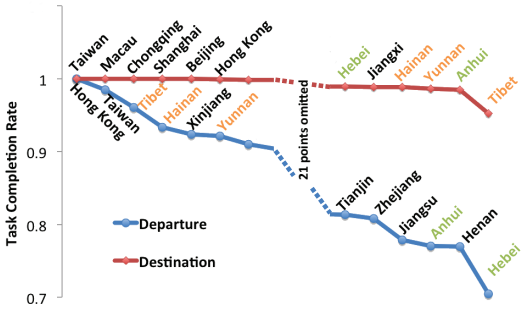


Figure 1: Task completion rate with respect to departure and destination provinces.

bust to the accents or dialects in those provinces. Note that, as our SDS will initialize the departure place according to user's GPS location if such information is available, most of those failed dialogs can still have their departure slots filled by default. Therefore, in the above figure, the task completion rates for departure provinces tend to be lower than those for destination provinces.

**Daily Time Periods** We also analyze the task completion rate of our system with respect to different daily time periods, as shown in Figure 2. It can be found that highest task completion rates occur during the midnight till early morning, and it decrease significantly in the evening, where the valley points are observed around 6pm and 8pm. It can be understood that people using the system in "abnormal" time periods (such as midnight to early morning) may have a strong requirement and motivation to have a journey booked. But in the evening (such as 8pm), one would expect that many users may just play with the app for entertaining purposes. A more attractive interaction strategy could be identifying those entertaining intentions and addressing them in a less formal manner. Environmental noise will be another factor affecting the task completion rate (e.g. the peak traffic hours 6pm∼7pm). A noise-level prior would improve the robustness of the SDS, particularly if a statistical system (Young et al., 2013) is employed in the future.

**User Responses vs. System Requests** Based on SLU-parsable utterances only, we investigate the correspondences between system requests and user responses, for which the results are illustrated in Figure 3. The statistics here aim to reflect user habits and to further examine the properness of the design of our system actions. It suggests that the users rarely say departure place and date in one

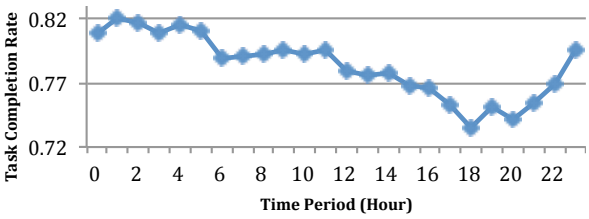


Figure 2: Task completion rate with respect to daily time periods.

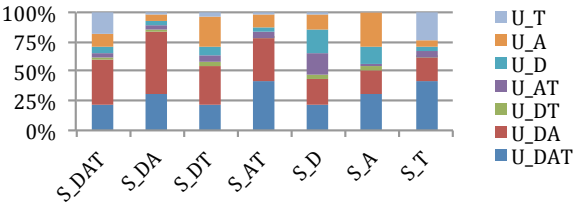


Figure 3: Correspondences between system requests (S) and user responses (U) for the three slots, departure (D), destination (A) and date (T).

utterance, therefore, the system may not sensibly benefit from asking for all such information simultaneously. Similar but slightly better correspondence is found for destination in conjunction with date as well.

## 3 Conclusion

This work discusses potential improvements to the granularity of SDS based on large-scale real user data analyses, for which the practical solutions will be the focus of our future work.

## Acknowledgements

The research in this paper is supported by the 973 Program No. 2014CB340505. The second author is supported in part by a SICSA PECE grant.

## References

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

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

J. D. Williams. 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.

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