A Prototype Spoken Natural Language Interface for Information Access on Mobile Phones *

Edward W.D. Whittaker, Paul R. Dixon, Josef R. Novak, Sadaoki Furui
(Tokyo Institute of Technology)

1 Introduction

In this paper we describe the motivation behind our prototype system for information access on mobile phones. The strategy that we adopt came out of research that was performed in several fields including question answering (QA) and automatic speech recognition (ASR). It also exploits some of the oddities of the Japanese mobile phone system, in particular the almost unrestricted ability to send e-mails from computers to mobile phones.

2 QA and ASR for Information Access

QA as a research topic has been studied for many years but it is only recently that practical applications have begun to be commercialized. While academia has continued to focus primarily on factoid, list and definition questions [2], commercial exploitation has not worried about this and instead produced systems that are not necessarily very complex but serve a useful function such as the sentence-based QA system Brainboost (`http://brainboost.com`).

Nonetheless, in contrast to keyword-based web-search QA has an inherent vulnerability since an answer is often be demonstrably right or wrong. Compared with the high expectation of good results (but additional subsequent effort for informational as opposed to navigational queries) that Google and others have accustomed its users to, QA systems suffer from the perception that they are sometimes very wrong which can lead to a lack of trust in other perfectly good answers. This problem is somewhat akin to the problems from which ASR has long suffered; users can easily point out the 'silly' (to a human) mistakes that dictation or dialog systems make and often ignore the fact that, in the long run, they save the user time, effort and probably money too.

Using ASR for information access in limited domains has recently accelerated with offerings from TellMe, Google and Medio that allow spoken keyword queries from mobile phones.

3 QA and ASR on Mobile Devices

Despite their imperfections, we believe that much of the potential at the intersection of these two important technologies can already be realised and is particularly well-suited to mobile devices especially mobile phones [1]. The motivation is simple: firstly, questions typically specify the information need much more precisely than a keyword query is able to so its use for information finding is well established, and secondly, questions are much more suited to spoken interfaces since speech is a more natural form of human communication than a keyword query is. Moreover, answers in response to questions are often not well suited to an audio channel since the user will typically want to do something

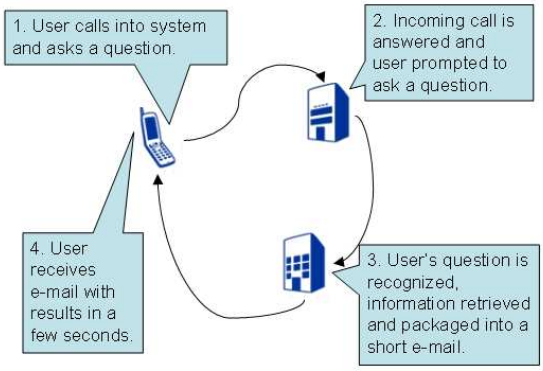

Fig. 1 High-level system flow diagram.

with part, or all, of an answer: store it for future reference, see the context in which it was given, forward it to a friend and so on. Consequently, we firmly believe that visual answers in the form of an e-mail, SMS, MMS or browser page, will almost always be preferred over a spoken answer. Depending on the scope of the questions that can be answered by a system we believe that the performance of ASR is already adequate for reliable and useful applications involving ASR and QA to be realised. In this setting mobile devices offer an almost perfect, ready-built platform for implementing such systems with their built-in microphone and display for multi-modal input and output of questions and answers.

A high-level diagram of the proposed system architecture is shown in Figure 1 and in Section 7 we give details on how to register and use our prototype system.

4 Domain Freedom vs. Freedom of Speech

Free-form spoken interaction with a machine has long been a part of science fiction and although we have recently come much closer to this goal we are still a very long way off. To build a useful open-domain spoken search system in the future will inevitably require some form of dialogue for clarification and expansion of the original query. In the meantime, building a successful system centres on the ability to minimise ambiguity in the task and there are several ways to approach this. While free-form speech input for open-domain QA is too difficult for the current state of the technology, free-form input for a train timetable system for example is perfectly feasible, assuming the user knows that the system they are calling can only provide answers related to train timetable queries and is unable to answer other types of queries. This allows us to be relatively flexible in terms of allowing different formu-

* 携帯電話による情報獲得のためのプロトタイプ自然音声インタフェース
エドワードウィッタッカー、ポールディクソン、ジョセフノバック、古井貞熙（東工大）

Fig. 2 Example system use and results.

lations of queries from the user and even permitting more specific information to be requested e.g. last train, next train, train arriving around 4.30pm etc. Alternatively, one might also envisage a system in which many domains are covered in which many different information requests can be supported. However, in this scenario the query formulation must be much more constrained for each individual domain if good performance is to be maintained. In both cases NLP techniques are applied to determine the user's precise information need from their question. Moreover, with the approach we are proposing, robustness to system errors can be built in, for example by providing N-best answers in the textual results. An analogous 'back-off' is very difficult to achieve when an audio channel is used exclusively.

5 Real Questions/answers vs. Academic Questions/answers

Open-domain factoid QA systems such as those used in annual international evaluations of QA technology are difficult to monetize for the simple reason that in reality there is not a huge demand for answers to factual questions. On the other hand, restricted domain QA systems that can answer questions covering more specific information are undoubtedly viable commercially. An oft-cited example is finding restaurants given a location and food type (see Figure 2), or train timetable information given starting and destination stations. In each of these scenraios there is a very wide range of possible questions that can be asked if the user is not first prompted for specific information. These might range from factual questions like *"Where can I get Dim Sum in Shinjuku?"* to procedural questions like *"How do I get from Tokyo Station to Shinjuku?"*. We believe that if the user were to speak the first question, the 'correct' answer should also include directions on where the restaurant is located and unasked for information such as opening times etc. since this must be considered a good answer although we are inferring the informational need of the user in such a case. If the user were to ask *"What time is the Taj Mahal in Shinjuku open until?"* the user would obviously like to know the specific closing time but will almost certainly also want confirmation that we have found the correct restaurant by providing an address as well, and maybe even competing possibilities if they exist. This represents a very marked divergence from the definition of a good answer in the TREC, CLEF and NTCIR evaluations.

6 Scope of Prototype and Performance

Our focus has been on restaurant-finding and train timetable applications, both staples of the ASR literature over the past twenty years or so but with few successful systems deployed to date. For restaurant-finding we employ a keyword spotting model with phone fillers that can recognise 1049 food-types and 1094 stations in the Tokyo area. This system uses results from GuruNavi as the answer provided to users. While not strictly question-answering, the speech interface does allow the user to formulate the question in a large number of ways yet still extract the relevant information. Currently, the system does not permit the asking of more complex questions with more specific information needs; this will be a focus of future research.

On a test set of 137 spoken utterances collected by the system from 7 users over a mobile phone connection, the system accurately recognised both the food and location 88% of the time, either the food and/or the location 96% of the time and got 4% of utterances completely wrong. Although the test set is too small to draw any firm conclusions this performance for a first-cut system with no language model is more than adequate to be useful for most users.

7 System registration and Access

To use the system a user must first register the telehone number and e-mail address of their mobile phone at `http://asked.jp/reg.html`. If the user has registered successfully they should receive a confirmation message on their phone. This confirmation message will inform the user what number to call to use the system. (The user must ensure that their phone is set to send the caller-id when a call is made.) When the user calls the number indicated they will be prompted to ask a question of a certain type (e.g. *"What kind of food do you want to eat and where?"*) and the user should speak their question in Japanese. After the query has been made, the user will be told to wait a few seconds for a message containing the results to be sent to the phone. At this point they can hang up.

8 Conclusion

In this paper we have described our motivation for using a multi-modal UI model for information access on mobile phones that consists of spoken questions as the input and textual answers as the output. We also described our prototype application and reported preliminary experimental results. Instructions on how to register and access the prototype system were also given.

References

[1] Harabagiu, S. and Moldovan, D., "Open-domain Voice-Activated Question Answering", *Proc. COLING* 2002.

[2] Voorhees, E. and Dang, H.T.,"Overview of the TREC 2005 Question Answering Track", *Proc. TREC2005*, 2005.