LANGUAGE MODELLING FOR THE IN-CAR INTELLIGENT PERSONAL ASSISTANT

Michael Schillo, Eric Atwell, Clive Souter, Tony Denson*
School of Computer Studies
University of Leeds
Leeds LS2 9JT
UK
schillo@scs.leeds.ac.uk
http://agora.leeds.ac.uk/schillo/

Abstract

This paper reports on research towards a natural language speech interface to a hand-held computer Personal Organiser. The research so far includes a review of related research and design of an experiment and analysis of the data. The application will be in-car, so that the user can update a computer database while driving. The focus now is on creating a natural sublanguage that suffices to access the computer data.

Keywords
sublanguage modelling, speech recognition, in-car, personal organiser, driving simulator

1. Introduction

This paper presents a pilot study of a potential industrial application of Natural Language Processing (NLP). An industrial sponsor, Visionair, are providing access to hardware and a practical application: the in- car "Intelligent Personal Assistant" system for travelling executives. The in-car application involves:

- a Psion hand-held computer "Personal Organiser", including diary, databases of addresses/appointments/etc.
- a portable PC-based commercial Speech Recognition system and development environment; currently we have a Speech Systems PE500 (described in (Churcher 95)), but other speech recognition systems are also available for evaluation and we envisage that a more powerful system will be needed for the longer-term application.

Our long-term aim is: to add a computer system that can access the information in the Psion personal organiser (including access to connected cellular telephone) via a Voice-User Interface (VUI), i.e. via a speech recognition system (such as PE500), so that it is possible to access any desired information via natural audio language input; furthermore, to extend this computer system beyond an interface, to include some of the "intelligence" and organisational functionality expected of a human Personal Assistant. Visionair are developing the application: software systems on the Psion which the VUI will access. This paper focuses specifically on the design of an English-like sublanguage for the in-car Personal Assistant.

Our approach involved a survey of potential users to elicit desired functionality, and collation and analysis of a corpus of datasets of samples of expected typical spoken interactions. Ideally we would like to allow an unconstrained range of possible utterances; but a fully-comprehensive language model that would cover all possible spoken inputs (e.g. spoken dialogues from the British National Corpus) would force the PE500's performance to be very low. So a central issue is the design of a Voice-User Interface (VUI) sublanguage, a subset of English which will allow users of a speech interface to issue commands in a reasonably 'natural' way while achieving acceptable speech recognition accuracy. The sublanguage should be tailored to cover the underlying functionality of the software system being interfaced to, but less constrained than a menu-keyword system: users should not be expected to have to learn a new language, but speak something near to "natural" English. This would simplify the syntax and increase the speech recogniser's accuracy and overall performance.

The fundamental base of such a model is a study of the possible utterances that occur. Having designed the VUI sublanguage, we will then seek an appropriate "off-the-shelf" speech recognition package to implement a prototype. Thus we view this as NLP rather than Speech Recognition research. Note that we have not included NL generation and speech output in our study. Visionair produce a mounting bracket to mount the Psion on the car dashboard, as an extra "dial", and the Psion screen is already used for some in- car output (e.g. the Find-It utility displays forthcoming landmarks, junctions etc. visually); other output can be directed to the in-car PC (e.g. diary updates, phone- dialling). At the moment, input is more of a practical problem than output; although we will consider speech output requirements later.

2. Analysis of the Psion functionality: what do we want to do?

We started by asking what a travelling executive might do with an Intelligent Personal Assistant (Schillo 95). No evidence of frequencies of certain functions has been gathered, and it has been found that there is no implicit constraint on the sequence of functions. Therefore Discourse Theory appears inappropriate (Churcher et al 96). In our setting similar problems with constraint come up as in (Atwell et al. 95) which deals with the air traffic control domain. Our survey began with current Psion users; we then consulted businessmen and executives who were not current Psion users but who expressed interest in a possible speech-driven alternative. We also consulted applications- providers (including Visionair) about current and prospective future functionality of the Psion. In the interviews, most people asked for, covers an agenda, to-do lists, memos, the cellular telephone and a navigation aid. Additionally, a range of so far utopic features have been asked for. Looking to the future some interviewees suggested the need to access other computer systems like Internet or the company's own computer resources.

3. Methodologies for corpus-based sublanguage design

As well as surveying desired generic functionality, we also needed a more concrete picture of what users would actually like to be able to say to access these functions. We surveyed research on corpus-based language modelling (e.g. (Thomas 93,96), (Gazdar 96)) and specifically on interfaces, and the design of elicitation experiments for Spoken Corpus datasets, as for example in (Heeman 95), (Isard 95) or (Pilkington 92). If we can identify the subset of language that is used to address the computer interface then we can improve speech recognition performance by restricting ourselves to just trying to recognise this subset. If we use the whole subset of language that is used to address the machine, we omit nothing necessary. To follow up this idea, we conducted an elicitation experiment. The experiment involved a task covering the whole functionality suggested by our survey, except for the navigation device (as this was expected to have a powerful functionality, beyond the scope of an initial prototype).

The recordings of these experiments will provide a guide to the sublanguage for the computer interface that should be sufficient to tackle all the possible queries to the computer.

Relevant to our work is evidence of convergence in queries between human users and a database system (Leiser 89) shows that by means of prompting, users can be 'lured' into using a smaller vocabulary and a smaller range of grammar, than they would usually do.

As our domain is a similarly constrained domain, we can expect that the same phenomenon applies in our case; therefore we need not collect a large corpus. We intend to use the convergence phenomenon in creating a more ergonomic system than would be possible otherwise.

4. Review of available resources

Having built a model of the relatively unconstrained language we would ideally like to use, we needed to survey the capabilities of current relevant technology. We looked at current in-car Voice-User Interfaces and other related speech-recognition systems, e.g. the Amerigon navigation system, which is virtually the only in-car NLP application that does not restrict itself to using a cellular phone. Amerigon created a navigation system with an isolated word, speaker dependent recogniser. The vocabulary appears to be under 50 words (digits 0-9, letters A-Z, and a few command words) with street names spelled out one letter at a time. The number of products on the market (as can be seen at the comp.speech newsgroup FAQ, see references) shows the great commercial interest on the field. But all these Speech Recognisers are still very poor in transcribing utterances of the whole set of possible utterances of a particular language in a natural environment. Additionally, it is not likely that a breakthrough will be reached in the near future.

5. Data collection

Emphasis was put on strictly objective data collection. During the design phase it was referred to literature from the field of Experimental Design (e.g. (McGuigan 93), (Miller 75)). The experiment confronted the Subjects with a series of tasks to be solved in co-operation with a human. This human should at the end of the project ideally be replaced by the targeted system. In the spirit of the objective data collection an icon language was invented so that Subjects would not be prompted by. Due to time and financial constraints we were only able to collect sixteen samples. It was however possible to collect these in two different settings. Namely a quiet office setting and the more noise but more natural setting of the driving simulator of the University of Leeds. The use of the Driving Simulator here is supporting the evidence that the data collected can be considered valid.

The design of the experiment allows us to continue collecting more samples, as soon as the findings in the analysis make it necessary. The samples so far, are thus considered a guideline for the project.

6. Analysis

In order to use the space provided here as efficient as possible, we would like to present the analysis of the data by an example.

One of the subtasks in the experiment was to make use of the cellular telephone. Every Subject encountered this subtask six times during their session, so we have a total of ninety-six utterances.

In order to get insight in the problem of reference during the dialogue one subtask was designed to make Subjects refer to their previous attempt to ring someone.

This gives us seventy-six utterances, all expressing the same function (one utterance is unclear, three were misinterpretations).

When comparing the utterances, one finds that they consist of forty-five different ways of saying. Thirty-two of those occur only once, which indicates that the number of samples is by far too small to be considered representative.

The phrase that occurred most often is:

"can you phone X please"

with 9 occurrences. An example for a curious way of addressing this function is for example:

"I need to phone X can you get him on the line please"

The interest of research is now in how easily can this amount of different phrases be condensed to a smaller model. Here a simple removal or rearrangement of the word "please" in the utterances deletes ten phrases (22.2 %)!

Similar findings are true for leaving slots for the verbs "to phone", "to call" and "to ring" instead of counting them as different phrases.

The second group of utterances that will be taken into account here, the group of utterances of the type "ring X again" first seem to be very divers. Analysing them after leaving a slot for

“again” or an “empty” slot deals with 5 utterances less (here 31.3%)!
Especially the relatively low usage of pronouns in this case is very encouraging (25%).
Similar data is available for all the subtasks performed during the experiment. As a preliminary conclusion we found that no fundamental difference in utterances recorded in the office and the driving simulator setting. This justifies making full use of the data collected in the office setting, despite it being apparently less “realistic” environment.

7. Feasible targets for the IPA

Provided a speech synthesiser (text to speech) and proper access from the PC system to the Psion data files are available, some of the following problems seem reasonably feasible. Under this assumption the functions will be accessible by "a kind of" menu structure: first the user states an utterance with a keyword indicating which application s/he wants to use and then use the constraints that are delivered by the functionality of the Psion. Judging from responses from Psion users, we believe that this would not be too inconvenient. Assuming that the following functions will be required:
- enter new memo
- read out a document specified by a filename or by string matching
- enter and ask for appointments
- add/modify to-do lists

we could use a mixture of menu traversing and free speech to increase the performance of a state-of-the-art speech recogniser. A request using this language model might look like this:

open my diary
what do you want to do with your diary?
read out my appointments for November 11th
-> the query is now being identified by the Speech Recogniser and the data will be returned.
(Taken from (Schillo 95))

If we use several levels, we gain performance, because we can enormously reduce the speech recognition language model on every level. However a big disadvantage of a hierarchical menu is that the user has to be constantly aware of his current position, as only “appropriate” language can be used. Hopefully we can withdraw as much as possible from the menu structure and lessen the disadvantages by using appropriate feedback from the system. A first draft of the structure is given in (Schillo 95).
Sequences of work were expected to have different probabilities in occurrence. This would have offered the possibility of doing research on probabilities and gaining constraints. Unfortunately we found no marked sequences or patterns in using different functions. However, we know that once a function is selected only a specific set of further requests is likely; this is evident by looking at the functionality.
We consider the following features not feasible in one year's time because they are too complex. However, we might cope with them later.
- depending on development of speech recognition, easier access to features
- being able to interrupt the system while it is reading out something
- changing your database, e.g. adding a phone number or reading out addresses specified by name (because the corpus is run-time dependent; one would have to implement a learning dictionary)
- create a model that adapts to the user, for example an ANN that anticipates the next function because of a internal statistics of the last sequences of work.

Other features have been discovered and are very interesting, but remain far too difficult to implement now. For more details see (Schillo 95).

8. Conclusions

The research so far has indicated that there is no need for unconstrained English recognition. Ideally we need to use a sublanguage less complex than the whole English language, but more complex than a menu-oriented command language. (Evans 95) reports similar conclusions for other restricted applications. However, certain subtasks will require a wider sublanguage, whereas others will be more constrained. Our findings in the above sections will guide us in pinning down what these constraints are. The results of the data collection experiment show that we can specify constraints for the recogniser. But our data also suggests that no sequences in using different functions will occur; so models from Discourse Theory cannot be applied.

9. Future Work

After further analysis of the samples, the next step will be evaluation of the sublanguage model with a current off-the-shelf speech recogniser. The sublanguage will than be improved using feedback from prototype tests (Schillo forthcoming).

References

(Atwell, et al. 95) Atwell, et al., *Developing a corpus-based grammar model within a commercial continuous speech recognition package* in Russell Collingham (ed.), Papers for the Institute of Acoustics Speech Group meeting on Integrating Speech Recognition and Natural Language Processing, Durham University. 1995

(Churcher et al. 96) Churcher, et al., *Using a commercial speech recogniser within the domain of Air Traffic Control*, Technical Report 96.04, School of Computer Studies, Leeds University 1996.

(Churcher 95) Churcher, Gavin E. *Current Speech Recognition Research System*. School of Computer Studies, Leeds, progress report, Nov 1995.

(Evans 95) Evans, Roger; Kilgarriff, Adam, *MRDs, Standards and How To Do Lexical Engineering*, Second Language Engineering Convention, pp. 125-33 95.

(Heeman 95) Heeman, Peter A. ;Allen, James, *The TRAINS 93 Dialogues*, TRAINS Technical Note 94-2, The University of Rochester, Computer Science Department, Rochester, New York 14267, March 1995.

(Gazdar 96) Gerald Gazdar (in press). *Paradigm Merger* To appear in Robin Milner & Ian Wand, eds., Research Directions in Computer Science, Cambridge: Cambridge University Press, 1996

(Isard 95) Isard, Amy; Carletta, Jean, *Transaction and Action Coding in the Map Task Corpus*, Human Communication Research Centre, University of Edinburgh 95.

(Leiser 89) Leiser, R.G., *Exploiting convergence to improve natural language understanding*, Interacting with Computers, Volume 1, pages 284-298, Butterworth & Co. 89.

(McGuigan 93) McGuigan, Frank, *Experimental Psychology: Methods of Research*, 1993.

(Miller 75) Miller, Steve, *Experimental Design and Statistics*, Methuen, pp. 117, 1975.

(Pilkington 92) Pilkington, Rachel M., *Question-answering for intelligent on-line help: The process of intelligent responding*, Cognitive Science, Volume 16 (4), pp. 455-491 92.

(Schillo 95) Schillo, Michael *What a travelling executive might do with an intelligent personal assistant*, School of Computer Studies, Leeds, progress report, Nov 1995.

(Schillo 96) Schillo, Michael *Report on related research*, School of Computer Studies, Leeds, progress report, Jan 1996.

(Thomas 96) Short and Jenny Thomas (eds.), *Using Corpora for Language Research: Studies in the Honour of Geoffrey Leech*, pp151-166,Longman, Harlow. 1996.

comp.speech frequently asked questions; Speech recognisers:
http://svr-www.eng.cam.ac.uk/
comp.speech/Section6/Q6.5.html

.

* Acknowledgement: We are grateful to Tony Denson and Visionair for sponsorship, inspiration and practical advise on this project.