OASIS  A FRAMEWORK FOR SPOKEN LANGUAGE CALL STEERING

Mike Edgington 1, David Attwater 2, Peter Durston 2

1 now at: SRI International
23 Millers Yard
CAMBRIDGE CB2 1RQ
United Kingdom
mike.edgington@cam.sri.com
http://www.speech.sri.com

2 BT Laboratories
MLB G pp4
IPSWICH IP5 3RE
United Kingdom
{david.attwater, peter.durston}@bt.com
http://www.labs.bt.com

ABSTRACT

OASIS is a research project at BT Labs investigating practical large-scale spoken language automation of call steering (call routing). BT's own operator assistance service is used as an initial trial domain. Spoken language call steering requires understanding of both user's language and behaviour. Therefore, the OASIS project makes extensive use of corpus-based machine learning techniques and iterative Wizard-of-Oz simulations to verify and refine the system design. We also describe a simple four layer model of general telephone call-handling, developed to characterise services in order to identify appropriate automation technology. We argue that for many helpdesk applications, a high proportion of callers will just describe the nature of their problem, rather than refer to particular services offered. For these calls, language use is highly unpredictable, and corpus based learning techniques are essential for speech recognition and understanding.

1. INTRODUCTION

In the past decade there has been a huge growth in call centres as companies increasingly rely on the telephone as the primary customer contact channel, resulting in the use of a single contact number for access to multiple services. In large organisations this has meant that telephone numbers alone do not uniquely define the final destination of a customer call, and has led to a proliferation of touch-tone menu-based call steering systems, which operate as a front end filter to the call centre. Unfortunately, these systems are difficult to use and can cause confusion.

This paper describes the OASIS project at BT Labs, which is investigating practical large-scale spoken language automation to tackle the call steering problem. A few research studies are also addressing the call steering task, namely the AT&T 'How May I Help You' project [1] and Lucent's call steering banking trials [2]. Section 2 introduces a general task based model of call handling. Section 3 describes a corpus of human-human calls to BT's operator assistance service, which is being used as a trial domain in OASIS. It also identifies several types of initial customer request, which can be defined in terms of the general call handling model. Section 4 describes the iterative Wizard-of-Oz experiments which are used in OASIS to verify system design, and describes a particular experiment investigating caller's response to various types of initial system prompt. Finally, Section 5 describes the current performance of the call classification algorithm.

2. CALL HANDLING MODEL

2.1 Caller Context

The context of a call to an automated service is very important. We note two important related dimensions:

- Victim or volunteer - was the caller expecting automation or were they unsuspecting victims?
- Frequent or infrequent - is the caller well primed and experienced or do they rarely call the service?

It is the clear experience of the authors that these two dimensions strongly dictate what can be achieved, in what style, for a given service. It should be noted that these two dimensions are not independent - callers are typically either infrequent victims or frequent volunteers, with occasional infrequent volunteers. In fact, frequent callers to a service quickly come to expect automation and become volunteers.

The term victim is deliberately emotive. In the United Kingdom IVR services, especially those based on touch tone, are widely disliked when callers are not expecting them [3]. Early indications are that acceptance of dialogue based speech recognition systems is higher, but there are currently no well established norms for talking with machines (see Section 4). Consequently spoken language behaviour from callers who have not been primed for a service can be difficult to predict.

2.2 Four Layer Model

There are typically four phases during a transaction with a service.

- Problem Specification – in which the problem to be solved is identified
• Task Identification – in which the customer intent is identified within the framework of available services
• Information Gathering – in which all details necessary to achieve the task are gathered from the customer
• Task Completion – in which the customer receives the service or information they require

In practice, when a customer calls a human agent there is often significant overlap between these various phases. For example there may be several stages of negotiation in order to discover the actual problem experienced by the customer, during which several potential services may be offered to the customer. Figure 1 shows a real call to a BT international operator, annotated into these four phases.

Operators Services, Jane Speaking
yes please I would like to make phonecall to Iraq Baghdad
phonecall to Iraq

Baghdad

OK er what you need to know the code do you
no I want - I want you to dial to Iraq Baghdad
you want me to dial for you
ya because from here it difficult do dial
right yes even for us it's difficult to
<laughing> dialling
right what's the number including country code please
er country code I think - 0 0 9 4
9 6 4
Baghdad N N N
N N N
dialling that for you

Task Identification

Problem Specification

Information Gathering

Task Completion

Figure 1: Call to BT international operator annotated with call handling model (operator turns in italics and customer turns in bold)

This model is helpful for analysing operator-based and automatic interactions. Note that the first two phases of a transaction may be implicitly satisfied. For example the BT directory enquiries (number listings) service on ‘192’ uses a human operator to achieve information gathering, and then automates the task completion phase by use of recorded number announcement. Since the directory enquiry service is very well known, the first two phases are implicitly fulfilled when the customer dials the ‘192’ access number.

3. OPERATOR ASSISTANCE CORPUS

BT has the largest call centre capability in Europe with around 115 call centres dealing with BT customer calls alone. These are staffed by the equivalent of nearly twenty thousand full time operators, taking over one billion calls per year. The most general contact point for BT is the operator assistance (OA) service, accessed through the well known ‘100’ code. Calls to OA operators cover an extremely wide range of topics, including: simple requests for information, malicious or inappropriate calls, explicit requests for non-OA services, requests for various BT services, and various miscellaneous calls such as highly ambiguous requests, confused customers, and the plain odd. Since the OA service has such broad functionality and customer profile, it was seen as a very challenging case study for helpdesk style dialogues.

3.1 Corpus Collection

The main OASIS corpus of human-human OA calls was collected in two phases. Phase 1 consisted of almost 1000 calls to the BT operator over a typical week. The calls were recorded onto DAT from an analogue connection to the operator headset. All calls were fully transcribed orthographically, including hesitations and restarts, and classified into detailed semantic classes. This database has been used in the pilot study to investigate general issues of dialogue structure, trends in language use and simple classification strategies.

The phase 2 data collection recorded around 25,000 calls over a month, and used a digital recorder attached to the OA switch, ensuring high quality recording and maximal separation of the operator and caller channels, making the recordings suitable for recognition experiments. This database is still in preparation.

3.2 Initial Customer Requests

By considering the detailed semantic classes, it was possible to characterise the customer’s initial response as one of four broad request types, which are summarised and exemplified in Table 1. This can be compared to the simpler taxonomy of short initial user utterances in a banking application described in [2].

Figure 2: Vocabulary growth of initial customer utterance

Figure 2 shows vocabulary growth as more calls are analysed. After 2500 calls, 2690 unique words were seen, and the vocabulary was still growing at 0.66 new words/call. Around half of these new words are proper nouns, but a substantial proportion are regular
common English words, (e.g. although, limited, regularly). These results are very similar to the AT&T study [1], although the vocabulary growth is somewhat higher in the OASIS corpus.

Table 1: Primary Request Types in Operator Assistance Calls

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Explicit service request (32%)</td>
<td>can I er, could you put me through to directory enquiries please</td>
</tr>
<tr>
<td>B</td>
<td>Implicit service request (17%)</td>
<td>can I have the number for the er Probation Office on Dover Street, S E 1</td>
</tr>
<tr>
<td>C</td>
<td>General problem specification (49%)</td>
<td>hi my name is XXXX XXXXXX and we have a problem here, there's someone whose trying to er call er who is calling us the whole time trying to fax us something, we haven't got a fax machine so must, must have the wrong number, it's been going on the whole day</td>
</tr>
<tr>
<td>D</td>
<td>Other (2%)</td>
<td>yeah I wanted them er to find out the how to spell a place that I wanted to send a telegraph to in Cornwall please</td>
</tr>
</tbody>
</table>

Figure 3: Initial customer utterance length for several call classes

The length of a customer’s initial utterance is an average of 18.3 words, with a median of 13 words. The useful range is from 1 word (‘hello’), through to 163 words (a caller having problems reaching the Italian tax office). However, the distribution of initial customer utterance length is partially dependent on the class of call, as shown in Figure 3. Reverse charge and alarm call are well-known service names and callers tend to request them succinctly. These are requests with explicit, well known service names (type A). Directory assistance calls usually take the form of a direct request for a particular number - directory assistance is rarely asked for explicitly. This is therefore an implicit request without use of a service name (type B). The line test class mostly represent problem specifications (Type C), so utterances are therefore more complicated and longer. Finally the 'other' category represents various requests including both short enquiries and much more involved problem specifications on different topics, resulting in a wide range of utterance lengths (mostly type C and D).

4. WIZARD-OF-OZ EXPERIMENTS

An ongoing series of Wizard-of-Oz (WoZ) experiments have been used to verify aspects of the system design, and to discover precise caller behaviour. Using the taxonomy described at the start of the paper, callers to the OA service are almost exclusively victim callers, so carefully controlled laboratory WoZ experiments would be inappropriate for this domain. Therefore all of the WoZ trials have been undertaken on live OA traffic, by connecting a computer simulation of the OASIS system directly into the OA centre, along with the digital call recording equipment used for the phase 2 corpus collection. This gives access to all of the callers speech for off-line speech recognition experimentation, alleviating the need for a live real-time recogniser in the field.

A fundamental assumption underlying the collection of the OASIS human-humans corpus is that callers will behave in a similar manner when they encounter an automated call steering system. A specific WoZ experiment was undertaken to verify this assumption.

4.1 Caller Behaviour at Initial Prompt

In order to investigate caller behaviour after the initial system prompt, several parameters were varied including the wording of the initial prompt and the formality of the speech style. A professional speaker was used to record all of the experimental prompts in several different styles of their choosing of varying formality. One of the prompts (Operator services, <name> speaking) is identical to that used by human operators, and acted as a calibration against the human-human performance.

Table 2: Proportion of uncooperative responses to initial prompt for different wording and speech style

<table>
<thead>
<tr>
<th>Operator services, &lt;name&gt; speaking</th>
<th>Hello, welcome to BT: How can I help you?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formal</td>
<td>52%</td>
</tr>
<tr>
<td>‘Chatty’</td>
<td>14%</td>
</tr>
</tbody>
</table>

Table 2 shows the proportion of calls which produced an ‘uncooperative’ first response in the caller for two prompt wordings with two tones of voice. An ‘uncooperative response’ was typically silence but also included the caller querying whether they had reached an operator, or exasperation at hearing an obviously automated system. As a control, analysis of the human-human corpus showed that approximately 10% of first responses were classed as uncooperative.

As can be seen from Table 2, the speech style of the prompt had an enormous effect on the initial caller response. Note that the second prompt wording, which is regarded by UK dialogue designers as more formal due to the word ‘welcome’, has slightly worse
performance than the more personal first prompt. Even in the very best case, there are more uncooperative responses than the 10% found in the human-human corpus, possibly indicating some residual indication of ‘professionalness’ in the recording. From this experiment, it is clear that the style and nuances of the initial recorded greeting require very careful attention to elicit co-operative responses. Research is ongoing to determine if the language use is still consistent with the human-human corpus.

This result has significant implications for the remainder of the dialogue in a call routing system. Firstly it suggests that the highest quality speech output is needed in these applications, which still requires the use of pre-recorded messages [4], limiting the flexibility of the dialogue. This result also suggests that callers do not have a model of how to respond to an automated system, yet from earlier work in the OASIS project it is clear that callers do want to know that they are talking to an automated system [3]. This is a central paradox in the design of natural dialogues for naïve callers. As speech recognition performance improves, speech output may become the limiting factor in holding highly natural interactions with customers.

5. CALL CLASSIFICATION

The OASIS project has used an information theoretic approach to classification using phrase fragments discovered from the training corpus. Results for OASIS phase 1 transcriptions are shown in Figure 4. This graph shows the confidence of the estimated call class for an accepted call, against the proportion of calls that should have been accepted but were incorrectly rejected (see [1] for a formal definition of these measurements).

![Figure 4: Text based classifier performance](image_url)

The graph shows both the recall performance of the classifier (tested on the training set) and the real classification performance, using an 80%-20% training-testing set split and cross-validated over 5 non-overlapping partitions of the corpus. The real classification performance is promising with around 93% recognition confidence at a false rejection rate of 30%. Note that the probability that the correct classification result is in the top-2 choices is over 97%. The interactive dialogue nature of the call steering task means that the second choice can be exploited using confirmation and disambiguation sub-dialogues [5].

5.1 Speech Recognition

![Figure 5: Simulated classifier performance 50% WER](image_url)

Results from recognition experiments are not available at the time of publishing, but state-of-the-art figures for conversational speech generally have high word error rates, around 40-60% [6]. The speech in the OASIS corpus contains large amounts of disfluency and initial indications are that error rates will be similarly high. Figure 5 shows the performance of the classification algorithm on a simple simulation of a 50% word error rate. (A random 50% of words in the transcriptions were subject to a substitution error, and replaced by a frequency weighted selection from the top 500 occurring words in the corpus.) At 35% false rejection rate, confidence in the first choice is down to 70%, but the top 2 confidence is just below 90%, which could still provide useful application performance. The classification performance with real recognition results is yet to be determined.

6. CONCLUSION

Automation of helpdesks requires careful design to encourage callers to engage. Even then, the nature of their enquiries will lead to unpredictable language use and corpus based learning techniques become essential.

7. REFERENCES