

# OASIS Natural Language Call Steering Trial

Peter J Durston<sup>1</sup>, Mark Farrell<sup>1</sup>, David Attwater<sup>1</sup>, James Allen<sup>1</sup>  
Hong-Kwang Jeff Kuo<sup>2</sup>, Mohamed Afify<sup>2</sup>, Eric Fosler-Lussier<sup>2</sup>, Chin-Hui Lee<sup>2</sup>

<sup>1</sup>BTexaCT, BT Adastral Park, Ipswich, IP5 3RE, UK.

{peter.durston, mark.farrell, david.attwater, james.3.allen}@bt.com

<sup>2</sup>Bell Labs, Lucent Technologies, 600 Mt. Ave., Murray Hill, NJ 07974, U.S.A.

{kuo, afify, fosler, chl}@research.bell-labs.com

## Abstract

A recent trial of natural language call steering on live UK calls to the operator is described along with its results. The characteristics of the problem are described along with the acoustic, language, semantic and dialogue modelling approaches employed. Natural language call steering is found to be viable, with recognition and semantic accuracy the current limiting factors.

## 1. Introduction

Agents in call centres for large companies are often segmented into different skill groups. This requires agents to identify relevant calls or steer them to another destination. Natural language call steering would therefore be advantageous.

Speaker independent spontaneous speech recognition over the telephone network is a difficult task. Topic identification for call steering however requires less information to be decoded from the speech and is becoming practical. Other research studies have addressed this problem, namely the AT&T 'How May I Help You' project [1] and Lucent's call steering banking trials [2].

This paper describes a recent trial of the OASIS call steering system on live traffic from a UK operator centre. This trial is part of an ongoing investigation under the OASIS project at BT laboratories, now BTexaCT, in collaboration with Lucent Bell Laboratories [3] [4].

In section 2 the OASIS corpus and the BT operator service are described. In sections 3 and 4 language modelling and acoustic modelling are discussed. Next, in section 5, the classification task is described, followed by the OASIS dialogue model in section 6. Finally the trial procedure and its results are presented in sections 7 and 8, and conclusions drawn in section 9.

## 2. OASIS corpus

The BT operator service handles calls on a wide range of topics from all over the UK. The service provides help with general connection difficulties and specific services such as transfer charge and reminder calls. It also receives a large number of calls that require re-direction to other BT call centres such as directory enquiries, faults or sales.

A corpus was collected by recording human-human operator service interactions. Each first utterance from the caller to the

operator was orthographically transcribed, and given a semantic and a dialogue move classification. Dialogue move boundaries were also coded. A subset of the calls were fully transcribed to support dialogue structure analysis.

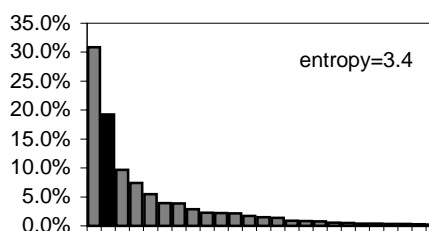


Figure 1. Semantic class distribution in the OASIS corpus. ('other' class shown in black)

The OASIS corpus contains first-turn customer utterances and is split into eight segments of 1000 utterances and a ninth of 441 utterances. Segment one was reserved as a test set.

In the version used for the trial there were 25 semantic classes and 6 dialogue move classes. The semantic classes represent commonly requested services and are distributed as in Figure 1. The 'other' class is marked in black and groups the unclassifiable sentences together. The most common class concerned connection problems. This class also contained the most varied language behaviour. The language characteristics for this task are described in [4] The move classes code the way in which the request was made and are similar to those discussed in the same paper. A study of operator services in the U. S. has shown similar language usage statistics but with a less skewed semantic class distribution [1][5].

## 3. Language modelling

The language model used for the trial was trained on all of the OASIS corpus excluding the held back test set.

The training set contained about 180K words and the test set about 25K words. The training corpus had a vocabulary size of about 4.5K words, but the language model vocabulary was restricted to those words that appeared at least twice in the training corpus. This restriction brought the vocabulary size down to about 2.5K. The out-of-vocabulary rate for the training set was about 2% and for the test sets about 3-4%.

The language model was augmented by adding phrases to the

original lexicon. A small subset of salient phrases used by the classifier for routing the calls was added to the lexicon for the language model. A total of 38 phrases were added that included greeting words like "hello-there," descriptions of the desired service like "wake-up-call," frequently used times like "one-o'clock," and special three digit UK service telephone numbers such as "1-5-4."

In addition to the phrases that were manually added, other phrases were automatically selected from the training corpus using an algorithm that is based on the maximum likelihood criterion [6]. Phrases were added iteratively if they improved the unigram likelihood of the language model with respect to the training corpus. Adding phrases can improve recognition results by capturing a longer context length for the language model. Examples of phrases added by this algorithm include "i-keep-getting," "i-want-to," and "constantly-engaged."

The trigram perplexity of the baseline language model with respect to the held-out test set was 43.2 and bigram perplexity was 54.4. Adding about 200 phrases reduced the normalised trigram perplexity by 2% and bigram perplexity by 20%. A relative improvement of 4% in recognition accuracy was observed with the trigram phrase based language model. This language model had about 2.4K unigrams, 39K bigrams and 92K trigrams. Syntactic clustering and generalisation of the training corpus, followed by generation of synthetic training sentences [7] improved the non-phrase language model by about 1-2% in word error rate, but preliminary results did not show any improvement for the phrase based language model. Mixing the trigram with bigram language models decreased the perplexity, but preliminary results did not show significant improvement in the recognition accuracy.

In addition to the language model that was used to handle the caller's first utterance, a language model that could handle confirmation and contradiction was also required. We artificially constructed a training corpus using phrase patterns including prefixes observed in smaller corpora from previous human-machine trials. These were populated with phrases from the OASIS corpus split at dialogue move boundaries to keep realistic syntactic contexts.

#### 4. Acoustic modelling

The baseline acoustic model that was used in the trial has previously been described [3]. This acoustic model was built using decision tree based state tying [8][9]. It is context dependent, including within and cross-word context dependent units. Due to the limited amount of in-domain data (around 12 hours at time of trial), out-of-domain data was also used to increase triphone coverage, and the acoustic models were adapted using the in-domain data [10].

In this paper we report results for this trial baseline and also subsequent improvements which led to a further 20% relative improvement in real-time recognition accuracy. Note that the trial results in section 8 of this paper do not reflect this improvement - thus we expect better overall system performance in future systems. We attribute the improvements in recognition accuracy to new pruning parameters, new acoustic models, and a new end-pointer. The acoustic modelling is based on triphones obtained from a

phonetic decision tree clustering. About 23 hours of in-domain telephone speech training data were used. About 438 hours out-of-domain UK English data was also available.

The ASR feature vector is a 38-dimension mel-LPC based cepstral vector, without energy component. To compensate for different channel characteristics, a real-time cepstral mean normalisation procedure was used. Different ways of utilising both in-domain and out-of-domain data were explored to improve the acoustic modelling component and hence the overall system performance. Three different methodologies were tried:

1. Mixing both in-domain and out-of-domain data for training. The amount of out-of-domain data was incrementally increased to find the best balance between the two types.
2. Adapting our baseline (trained using all the in-domain and out-of-domain data) using in-domain data. We tried several iterations of MAP [11], and SMAP[12] with different subsets of in-domain data.
3. Using only in-domain data.

Interestingly, it was found that for this task using only the in-domain data led to comparable results with the other methodologies while producing much smaller models. In addition, for the adaptation experiments several iterations were required to be able to reach the performance of the in-domain model. This suggests that the out-of-domain data is not characteristic of this task.

Speech end-pointing for real-time decoding was also very important. By incorporating a new real-time end-pointer [13], very similar recognition performance to batch experiments was achieved. Without the new end-pointer, about 5% in recognition accuracy would have been lost.

Many experiments were run in order to determine the optimal values for the pruning parameters. The wrong tradeoffs to achieve real-time would result in poor performance. After all the improvements were made, the real-time recognition performance went from 49.3% word correct (WCorr) and 59.4% word error rate (WER) to 58.6% WCorr (48.1% WER). The best non-real-time results achieved so far is 64.3% WCorr (42.2% WER). In the trial a real-time accuracy of 49.5% WCorr (62.1% WER) was achieved. After the recogniser was optimised as described above, these recognition results improved to 57.7% WCorr (50.6% WER).

The word error rates are high because of many possible reasons. Some of these include insufficiently well-labelled and relevant training data, adverse acoustic conditions, multiple regional accents in the UK, fast speaking rates with significant phone deletions, and disfluencies. It is likely that with additional in-domain acoustic and language data, the word error rate can be further reduced.

#### 5. Semantic modelling

The OASIS system used a fragment-based semantic classifier using salient phrases as described in [3] and similar to [1]. Vector-based information retrieval techniques have also been shown to be effective for this domain [17]. A simple parser

was also used prior to classification to detect significant features in the recognition output such as telephone numbers, times, cities and user preferences for terminology. These features allow some generalisation in the classification and also provide additional information to the dialogue manager.

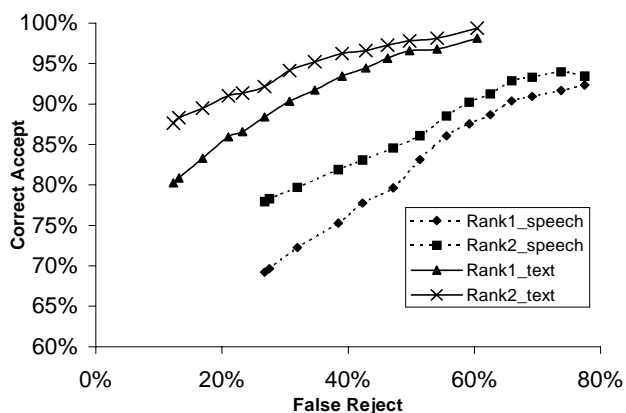


Figure 2. Classification accuracy for text and speech.

Figure 2 shows the accuracy of the semantic classifier tested on the OASIS database with text and also with speech recognition results reported using the conventions described in [1]. Recall that calls correctly classified as 'other' at rank 1 are considered correctly rejected.

## 6. Dialogue modelling

The OASIS dialogue model is a finite state machine with one state for each logical question-answer turn-pair. The dialogue design is described as a network of states emanating from a single start state. Arcs in the network are directional and ordered. Each arc has a single boolean condition assigned to it. These conditions refer to the contents of an enquiry blackboard which contains sets of feature-value pairs with associated confidences. Transitions between states are decided by fully parsing this network depth-first from the start state on completion of each recognition, classification and blackboard update. The first state that is found to have a true condition entering it and no true conditions leaving it is chosen as the next dialogue state. Re-parsing for every turn permits full mixed-initiative dialogue to occur allowing the possibility of transition between any two states at every turn. More traditional finite state dialogue systems, for example the W3C VoiceXML, may also be considered a sub-set of this scheme allowing them to be integrated without structural change [14].

Prompting in the OASIS system is via concatenation of recorded prompts. Textual phrases may also be used if text to speech is available. A dialogue state selects what is termed a logical prompt depending on dialogue state history - also maintained on the blackboard. Logical prompts are simply non-terminals or terminals in a generative grammar of speech files or textual phrases. These are then used to generate the actual terminal sequence of prompts to play given a particular blackboard state. Often a logical prompt will be a single recorded speech file but can also be complex sequences of prompts - for example to generate a time of day or a date. The choice of a logical prompt for a given dialogue state is

described by a set of dialogue state history N-grams - one or more for each logical prompt. By default at least one logical prompt must be prescribed for each dialogue state i.e. there must be a full set of unigrams. The prompting subtlety may then be extended arbitrarily by adding further N-grams to adjust prompting for more diverse dialogue contexts.

Recognition grammars are also selected according to dialogue state. States may share a grammar and grammars may be context-free or N-gram based.

The dialogue model also permits inference rules to be specified as condition/action pairs acting on the blackboard whenever it is updated. Thus interrelationships between classes and features may be arbitrarily represented, for example supporting relational approaches such as [14] or [15].

## 7. Trial Procedure

The trial system used the Lucent Bell Laboratories research recogniser on their STIP platform[16]. Dialogue management and classification components were supplied by BTexaCT.

The trial system took 725 live calls to the Leicester BT operator centre over two days in October 2000 during day time working hours. As usual callers experienced a queue length typically around 15 seconds. They then heard a "Please wait" in a male voice and were diverted to the trial platform. This introduced a delay of 5 seconds ending in a single ring tone. Callers were then greeted by a female voice in a conversational style with "Hello! this is the automatic operator. How can I help you?".

This trial methodology meant that callers were genuinely motivated and calling with real requests under real network and environmental conditions. Callers had no expectation of deviation from the normal manual service.

## 8. Trial Results

### 8.1. Initial engagement

Culturally UK callers show considerable resistance to any form of automation. Engagement is therefore an important factor. For this reason, following the initial prompt callers who remained silent or who said "hello?" or similar were re-greeted. This has been found to be an effective engagement strategy.

A total of 9% of callers hung up at some stage during the engagement dialogue. This was expected as previous trials have observed similar or higher rates for any unexpected changes to the service. Only 4% of callers defaulted to the operator without engaging meaningfully with the machine. The role of prompt style in this engagement is discussed in [4].

### 8.2. First engaged utterance performance

Utterances from callers which were not classified as silence or a greeting were termed the *first engaged utterance*. Figure 3 shows the different proportions of classification outcomes for these first engaged utterances along with confidence bands. In the dialogue rejected utterances were transferred to the

operator. Low confidence acceptance led to explicit confirmation. High confidence acceptance led to immediate fulfilment or call steer.

Outcome	Proportion
Correct Accept(high)	5.3%
Correct Accept(low)	35.5%
Correct Reject	13.1%
False Accept(high)	0.7%
False Accept(low)	27.9%
False Reject	17.5%
Total	100.00%

Figure 3. Semantic classification accuracy for first engaged utterance. Confidence shown in parentheses.

A total of 53.9% of first engaged utterances were correctly classified. Only 0.7% were given the wrong service fulfilment without the opportunity to contradict.

### 8.3. Performance on correct confirmation

As seen above 35.5% of first engaged utterances led to an explicit confirmation of the correct service. In response to this, 59% of these cases were confirmed and led to a successful outcome. The system was designed to be fail-safe and any uncertainty in confirmation was rejected. This led to a large number, 35%, defaulting to the operator at this stage. Most of the remaining 6% hung-up probably due to dialogue wording issues.

### 8.4. Performance on incorrect confirmation

Also seen above 27.9% of first engaged utterances led to an explicit confirmation of the incorrect service. Of these 22% were corrected successfully at the confirmation stage, 12% hung up and 8% deliberately default to the operator. 7% were incorrectly steered mostly due to an incorrect 'yes' response from the caller not listening to, or misunderstanding, the prompt. The remaining 51% defaulted to the operator representing missed steering opportunities.

## 9. Conclusion

The recognition task for UK spontaneous English is difficult and callers in the UK have little motivation to engage with machines. In spite of this, the trial has shown the viability of natural language call steering with very low false steer rates. This was achieved at the cost of relatively high proportion of confirmations. When coupled with the rich diversity of behaviour following wrong confirmations this suggests a greater need to get the class right first-time. Thus recognition and classification accuracy, and the coupling of the two, are seen to be the current limiting factor and are fruitful areas for further research. Having said this, incorporating the latest improvements in recognition accuracy, it is expected that almost half of all callers to this service may be steered correctly without human intervention.

## 10. Acknowledgements

Grateful thanks are extended to other members of the team at Bell Laboratories - Wu Chou, Qiru Zhou, and Antoine Saad who contributed strongly to the early stages of the project. Thanks also to Jinsong Zheng, Peter Li, and Olivier Siohan for

help with the core recogniser. Thanks to BT Retail who gave permission for the trial results to be published.

## 11. References

- [1] A Gorin, G Riccardi, J Wright, "How may I help you?" *Speech Communications* 23 (1997), pp. 113-127.
- [2] J. Chu-Carroll and B. Carpenter, "Vector-based natural language call routing," *Computational Linguistics*, vol. 25, no. 3, pp. 361—388, 1999.
- [3] Wu Chou, David Attwater et al; "Natural Language Call Steering for Service Applications" in *Proc. ICSLP*, (Beijing, China), Oct. 2000.
- [4] M. Edgington, D. J. Attwater and P. J. Durston, "OASIS - a framework for Spoken Language Call Steering", in *Proc Eurospeech'99*.
- [5] A Gorin, J Wright et al. "Semantic Information Processing Of Spoken Language," in *Proc. of the Workshop on Multilingual Speech Communication*, ATR Laboratories (Kyoto), Oct 2000.
- [6] H.-K. J. Kuo and W. Reichl, "Phrase Based Language Models for Speech Recognition", in *Proc. EuroSpeech'99*, (Budapest, Hungary), Sept. 1999.
- [7] E. Fosler-Lussier and H.-K. J. Kuo, "Using Semantic Class Information for Rapid Development of Language Models Within ASR Dialogue Systems," in *Proc. ICASSP'01*, (Salt Lake City, Utah), 2001.
- [8] W. Chou, "Decision Tree Tying Based on Penalized Bayesian Information Criterion," *Proc. ICASSP'99*.
- [9] W. Reichl and W. Chou, "Robust Decision Tree State Tying for Continuous Speech Recognition", *IEEE Trans. on Speech and Audio Processing*, Jan. 2001.
- [10] W. Chou, O. Siohan, T. Andr'e Myrvoll and C.-H. Lee, "Extended Maximum A Posterior Linear Regression (EMAPLR) Model Adaptation for Speech Recognition," *Proc. ICSLP'2000*, Beijing.
- [11] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation of multivariate gaussian mixture observations of Markov chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 291--298, April 1994.
- [12] K. Shinoda, and C.-H. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Trans. on Speech and Audio Processing*, vol. 9, pp. 276--287, March 2001.
- [13] Q. Li, J. Zheng, Q. Zhou, and C.-H. Lee, "A Robust, Real-Time Endpoint Detector with Energy Normalization for ASR in Adverse Environments," in *Proc. ICASSP'01*, (Salt Lake City, Utah), 2001.
- [14] D Attwater, J Fisher, H Greenhow. "Towards fluency - structured dialogues with natural speech input." in *BT Technology Journal*, pp178-186. Vol. 17 No. 1. 1999
- [15] J Wright, A Gorin, A Abella. "Spoken language understanding within dialogs using a graphical model of task structure" in *Proc. ICSLP*, Paper 385, Sydney. 1998.
- [16] Q. Zhou, C.-H. Lee, W. Chou, A. Pargellis; "Speech Technology Integration and Research Platform: A System Study," in *Proc. EuroSpeech'97*, (Rhodes, Greece), Sept 1997.
- [17] H-J K Kuo and C H Lee. "A portability study on natural language call steering" in *Proc. Eurospeech*, (Denmark), Sept 2001.