Proceedings of the Institute of Acoustics

SEMANTIC INFORMATION PROCESSING OF SPOKEN LANGUAGE[1]

A.L. Gorin, J.H. Wright, G. Riccardi, A. Abella and T. Alonso

AT&T Labs, Speech Research
180 Park Avenue
Florham Park, N.J. 07932
{algor, jwright, dsp3, abella, tma}@research.att.com

ABSTRACT

The next generation of voice-based user interface technology will enable easy-to-use automation of new and existing communication services. A critical issue is to move away from highly-structured menus to a more natural human-machine paradigm. In recent years, we have developed algorithms which learn to extract meaning from fluent speech via automatic acquisition and exploitation of salient words, phrases and grammar fragments from a corpus. These methods have been previously applied to the *'How may I help you?'* task for automated operator services, in English, Spanish and Japanese. In this paper, we report on a *new* application of these language acquisition methods to a more complex customer care task. We report on empirical comparisons which quantify the increased linguistic and semantic complexity over the previous domain. Experimental results on call-type classification will be reported for this new corpus of 10K utterances from live customer traffic.

1. INTRODUCTION

The next generation of voice-based user interface technology will enable easy-to-use automation of new and existing communication services. A critical issue is to move towards a more natural human-machine paradigm. By *natural*, we mean that the machine understands what people actually say, in contrast to what a system designer would like them to say. This approach is in contrast with menu-driven or strongly-prompted systems, where many users are unable or unwilling to navigate such highly structured interactions. This research targets shifting the burden from human to machine, wherein the system adapts to peoples' language, as contrasted with forcing users to learn the machine's jargon.

In particular, we have developed algorithms which learn to automatically extract meaning from fluent speech. A key intuition is that some linguistic events are crucial to recognize and understand for a task, others not so. We've quantified this idea via *salience*, which measures the information content of an event for a task [Go95]. Algorithms have been developed which automatically acquire and exploit salient words, phrases and grammar fragments from a corpus [Go97][Wr97][Ar99]. These methods have been previously applied to the *'How may I help you?'* task for automated operator services, in English [Go97], Spanish and Japanese [Ba00]. The early experiments were based on excerpts from human/human interactions drawn from live customer traffic. In later experiments, spoken language understanding (SLU) was then embedded in a dialog system [Ab97][Ab99] and experimentally evaluated [Ri00] on 20K human/machine transactions, again drawn from live customer traffic.

The primary focus of SLU in these experiments has been call-type classification, i.e. determining which service type a customer is requesting. Other researchers have reported on analogous experiments in other domains [Ca98][Ed99]. In the operator services domain, we've also developed methods for extracting auxiliary information such as phone and credit-card numbers embedded in natural spoken language [Ra99].

[1] This paper previously presented at the Workshop on Multilingual Speech Communication, Kyoto (Oct. 2000).

In the operator services domain, the task involves placing telephone calls, specifying billing methods for those calls (e.g. collect, card, etc.), and requesting information about making those calls (e.g. rate, area codes, etc). In this paper, we report on a new application of our language acquisition methods to a more complex customer care task. In this task, users are asking questions about their bill, calling-plans, etc. This is intuitively a more complex domain.

In this paper, we report on empirical comparisons which quantify the increased linguistic and semantic complexity of this new task over the previous domain. Experimental results will be reported and compared for a new corpus of 10K human/human dialogs recorded from live customer traffic. In Section 2, we describe the new database and how it was collected. Section 3 discusses the semantic complexity of this customer care task and compares it to the operator services domain. In Section 4, we do the same for linguistic complexity. An initial experimental evaluation of call-classification from speech for this customer care task is reported in Section 5, demonstrating portability and scalability of our language acquisition methods.

2. Database

This 'initial data collection' for the customer care task was extracted from recordings of customer interactions with human agents. The first few minutes of 10K transactions were recorded directly in digital mu-law format. From each transaction, the first 'task-oriented' customer utterance was manually segmented. This utterance was then transcribed and labeled. From these, 8K utterances were randomly selected as a training set and 1K for testing. An example of a request for an account balance is as follows:

"[uh] I need to check how much I owe [.brth] I apparently didn't mail a payment again I want to see what [uh] what you have there"

While these utterances are extracted from human/human interactions, we are of course eventually interested in human/machine interactions. For the operator services domain, we similarly started with human/human interactions in [Go97] as a precursor to constructing an automated spoken dialog system. From that system we then collected the human/machine interactions analyzed in [Ri00]. In this paper, when comparing complexities between the two domains, we will of course use the human/human utterances from the operator services task.

Intuitively, one expects users' language to simplify when talking with a machine. This expectation was validated and quantified in the comparisons between spoken language in human/human and human/machine transaction in the operator services domain [Ri00].

3. Semantic Complexity

We recall that in the operator services task, there were 15 call-types [Go97] with auxiliary information such as phone and card numbers [Ra99]. Initially, it was convenient to view these call-types as an unstructured list. and the task simply as classification [Go97]. It became clear, however, that there was additional structure and relationships amongst these labels. For example, collect is *a kind of* billing method. In addition, any call *has a* forward-number (the number being called), and a card call furthermore *has a* card-number to be billed. We then quantified these *'is-a'* and *'has-a'* relationships in an objected oriented *inheritance hierarchy* [Ab97][Ab99] for the dialog manager. This inheritance hierarchy was reflected in the task structure graph of [Wr98], so that the original classification problem evolved into mapping an utterance onto a probability distribution over a graph.

In the customer care task, we defined 19 call-types and 12 auxiliary elements. There are a multitude of *'is-a'* and *'has-a'* relationships amongst these labels, many more than in the operator services domain. For example, the following utterance would be labeled as a general billing query:

"I have a question about my bill"

Consider then the account balance example given previously, which *is a kind of* billing query. An example of a query regarding unrecognized numbers is as follows:

" [um] I was just calling in regards to some [uh] couple of phone numbers that I do not recognize"

Such an unrecognized-number query is *a kind of* question about a *charge on the bill.* Given the complexity of a bill, there are furthermore many auxiliary elements which are associated to any particular item or charge. An item being questioned *has a* date, item number, page number, etc.

Although we do not yet have a method of quantifying the complexity of such inheritance hierarchies, we do observe that the number of nodes and arcs is more than doubled as compared to the operator services task.

A second dimension of semantic complexity is multiplicity of labels. This was a rare phenomena in operator services, comprising less than 2% of the utterances. In customer care, 12% of the utterances involve multiple call-types, a 6-fold increase. A third dimension is the open set component of the task. As discussed in [Go97], in any real-world task, there are always utterances which do not match any of the defined call-types, which we then denote *other* and need to be routed to a human agent for handling. A crucial feature of the SLU is thus rejection, i.e. knowing what it knows. In the operator services task, *other* comprised 12% of the utterances, while in customer care it is 26% using the current label set. I.e., more than twice the open-set component.

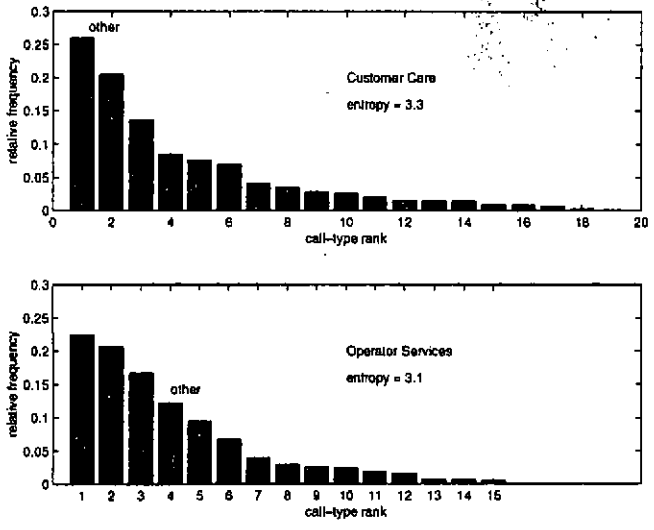

Figure 1. Semantic Rank-Frequency Distribution

Given the 19 call type labels for this task, we can measure the skewness of their distribution. In Figure 1, we plot the rank frequency distribution of these 19 call types, which have an entropy of 3.3. For comparison, we include the corresponding distribution for the 15 call-types from operator services [Go97], which has an entropy of 3.1.

4. Linguistic Complexity

We now address the linguistic complexity of these two tasks. Not surprisingly, these measures reflect the increased semantic complexity of the task. In Figure 2, we plot the number of words per utterance in the two tasks, observing that the customer care utterances are much longer than the operator services task. In particular, the average number of words per utterance for customer care is 39, more than double that of operator services which was 19 [Go97].

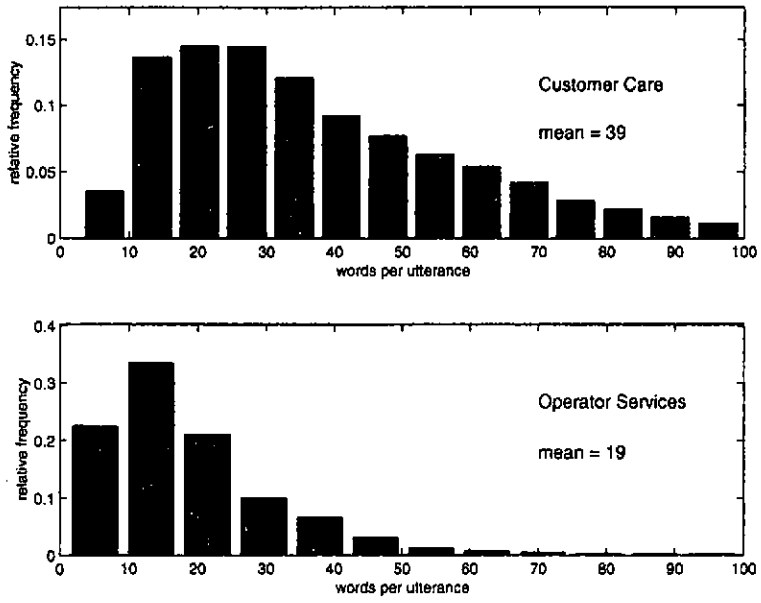

Figure 2. Utterance Length Distribution

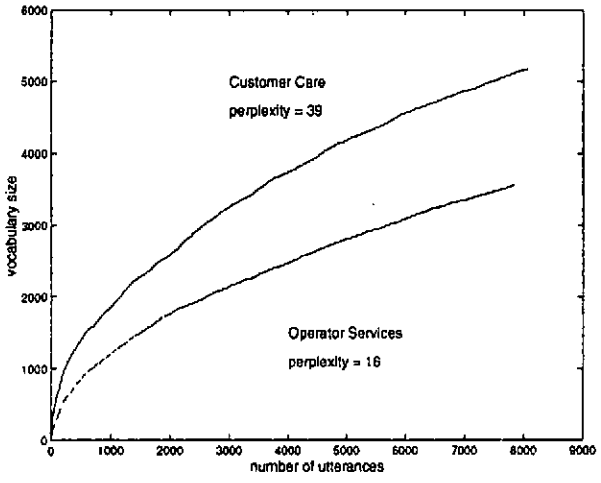

Figure 3. Vocabulary Growth

Next, we measure the vocabulary growth for the 8K training utterances in both tasks, as shown in Figure 3. The vocabulary size after 8K utterances from customer care is 5.2K, compared to only 3.6K for operator services. This is not surprising, given the longer utterances and greater semantic complexity. We also observe that the OOV rate (per utterance) is similar (as seen from the slopes of the vocabulary plots). We then computed the test set perplexity of each using VNSA phrase-bigram language models [Ri96], observing that the perplexity of the customer care language is 39, more than double that of operator services, which was 16 [Go97].

5. Call Classification from Speech

We now report on baseline experimental results evaluating call classification from speech on this new task. All components of training are on the 8K training utterances, and testing is on a separate 1K test-set.

First, off-the-shelf telephony acoustic models were adapted with this training data. Then, VNSA phrase-bigram language models [Ri96], salient grammar fragments [Wr98] and a classifier were then trained. These salient fragments are then detected in the ASR output and exploited for classification [Go97]. In Figure 4, we plot the number of salient fragments detected per utterance for both tasks. Observe that although the utterances are twice as long, there are less than half as many salient fragments detected in customer care versus operator services.

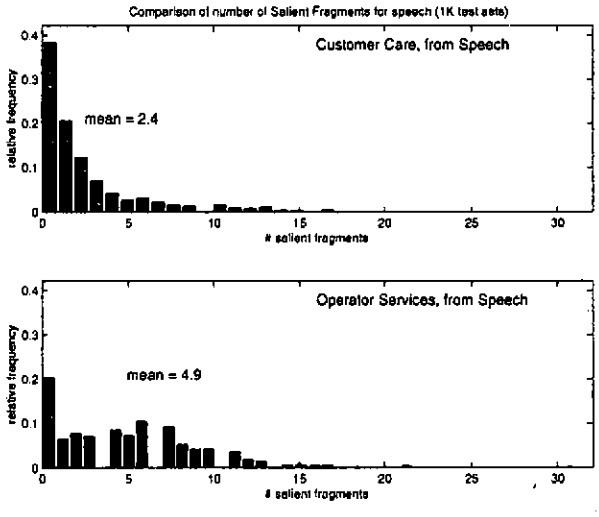

Figure 4. Number of Salient Fragments Detected per Utterance

In Figure 5, we plot the percentage of each utterance which is covered by salient fragments. We observe that the coverage in these longer customer care utterances is only 13%, less than half of the 31% coverage in operator services.

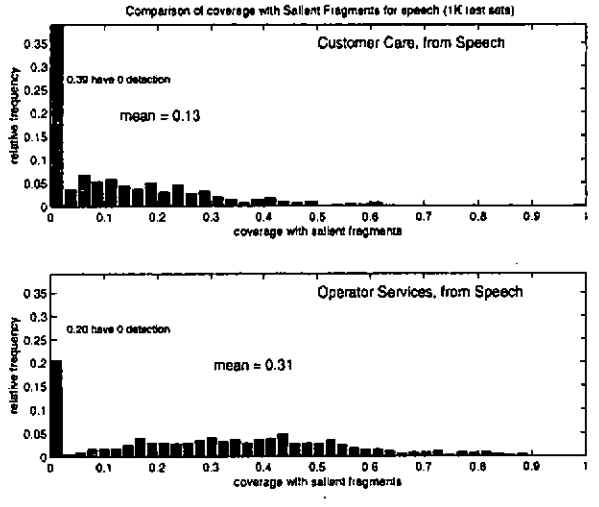

Figure 5. Coverage of Utterances by Salient Fragments

Finally, in Figure 6, we plot the ROC for call-classification on the customer care task. Recall [Go97] that False Reject Rate FRR) is the percentage of rejected utterances which were labeled as one of the non-other call-types. Such rejected calls are routed to a human agent, so the cost of such an error is a missed opportunity for automation. FRR can be traded off against Probability Correct by varying the rejection threshold. Results are shown for rank 1 and rank 2 from both text and speech.

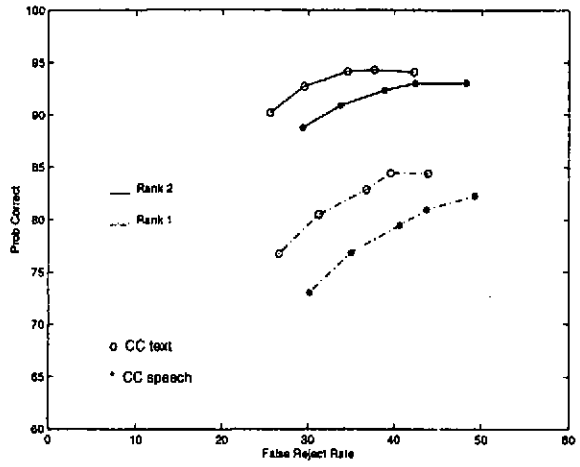

Figure 6. Call Classification Performance

6. Conclusions

Although the entropy and number of call-types in the two tasks are similar, there are still several dimensions in which customer care is more complex: number of auxiliary elements, number of relationships amongst labels, multiplicity of labels, and open set component. Linguistic complexity is more than double in several dimensions. Remarkably, given the increased complexity and sparse coverage, one can still achieve over 90% correct classification from speech, albeit with a FRR of 30%.

7. REFERENCES

[Ab97] A. Abella and A.L. Gorin, "*Generating Semantically Consistent Inputs to a Dialog Manager*," Proc. Eurospeech, Greece, pp. 1879-1882, Sept. 1997.

[Ab99] A. Abella and A.L. Gorin, "*Construct Algebra: Analytical Dialog Management,*" Proc. of the ACL, Washington D.C. June 1999.

[Ar99] A. Arai, J. Wright, G. Riccardi and A. Gorin, "*Grammar fragment acquisition using syntactic and semantic clustering*," Speech Communication, vol. 27, no. 1, Jan. 1999.

[Ba00] S. Bangalore and G. Riccardi, "*Stochastic Finite-State Models for Spoken Language Machine Translation*", Proc. Workshop on Embedded Machine Translation Systems, NAACL, pp. 52-59, Seattle, May, 2000.

[Ca98] R. Carpenter and J. Chu-Carrol, "*Natural Language Call Routing: A Robust Self-Organizing ApproachProceedings of Eurospeech 99. Budapest, Hungary,*. pp. 76-79, Sydney (1998).

[Ed99] M. Edgington, D. Attwater, P. Durston, "*OASIS - a Framework for Spoken Language Call Steering,*" vol. 2, Page 923-926," Proc. ICSLP, Sydney (1998).

[Go95] A.L. Gorin, "*On Automated Language Acquisition*," 97(6), pp. 3441-3461, Journal of the Acoustical Society of America (JASA) (June 1995).

[Go97] A.L. Gorin, G. Riccardi and J.H. Wright, "*How may I Help You?*", Speech Communication 23 (1997) pp. 113-127.

[Ra99] Rahim, M., Riccardi, G., Wright, J., Buntschuh, B. and Gorin, A., "*Robust automatic speech recognition in a natural spoken dialog,*" Workshop on Robust Methods for Speech Recognition, Tampere, Finland, 1999.

[Ri96] G. Riccardi, R. Pieraccini and E. Bocchieri, " *Stochastic Automata for Language Modeling* ", Computer Speech and Language, vol. 10(4), pp. 265-293, 1996.

[Ri00] G. Riccardi and A.L. Gorin, "*Spoken Language Adaptation over Time and State in a Natural Spoken Dialog System*," IEEE Trans. on Speech and Audio, Jan. 2000.

[Wr97] J.H. Wright, A.L. Gorin and G. Riccardi, "*Automatic Acquisition of Salient Grammar Fragments for Call-Type Classification*", Proc. Eurospeech, Greece, pp. 1419-1422, Sept. 1997.

[Wr98] J.H. Wright, A.L. Gorin and A. Abella, "Spoken Language Understanding within Dialog using a Graphical Model of Task Structure," Proc. ICSLP, Sydney, 1998.