"In a moment we will give you instructions on how to enable call blocking on your cell phone. We will also send these instructions by text message and to the email address on file. You can hang-up now, or hold for spoken instructions. {Pause 3s}. To enable call blocking you need to follow three simple steps. First ... "

As cell devices become increasingly integrated for voice and data then this will increase dramatically with the real promise of useful multi-modal interaction in the future. Imagine:

- Pushing menus onto the screen dynamically during a call
- Pushing instructions onto the screen during the call.
- Interactively interrogating the cell phone for its current location and advice relevant to the current location.

The possibilities are endless...

David Suendermann

david@speechcycle.com

Let Data Rule: Context-Adaptive Statistical Grammars

Abstract

Only recently, we showed that using very large amounts of transcribed and annotated utterances collected in spoken dialog systems to replace rule-based by statistical grammars may produce a significant performance gain. However, as more and more data is collected, one can start specializing beyond one-by-one grammar replacement. We claim that the availability of huge amounts of call log, transcription, and annotation data can be exploited to produce *context-adaptive statistical grammars* which optimally fit every possible state in a dialog system.

The Problem

Large-scale exploitation of transcription and annotation of all the utterances collected by a speech recognizer enabled us to extensively replace handcrafted rule-based grammars with statistical grammars (Suendermann et al., 2009). We also observed that even a very small number of utterances (1000 or less) are enough to train statistical grammars which consistently outperform their rule-based counterparts. However, when higher numbers of utterances are being used the relative performance gain considerably flattens. Moreover, we observed that when grammars are used in different dialog states they exhibit low performance in some individual contexts, while their average performance is high. For instance, yes/no grammars are trained on hundreds of thousands of utterances collected over hundreds of recognition contexts of multiple dialog systems. In the application at hand, the overall number of affirmative answers is higher than negative responses when computed across all contexts. Thus, a statistical context-independent grammar assigns a high a-priori probability to “yes” and a low probability to “no” and their synonyms. However, the grammar would perform poorly when applied to a specific context where the number of negative responses is significantly higher than the positive ones. There are many variables, besides the recognition context, that may affect the

distribution of user inputs, including differences in the caller population, individual user profiles, external events like outages or marketing campaigns, time of day, day of the week, or the history of the interaction.

Utilizing the notion of state-based dialog management (Minker and Bennacef, 2004), we observe that the user input distributions are state-dependent causing grammar performance to vary accordingly.

The Solution

A reasonable way to overcome the context dependency problem would be that of training specific context-dependent grammars. However, one has to consider the trade-off resulting from using more data for a single grammar vs. using less data for each individual one: For instance, by using all data available in all contexts for a single general grammar, one may obtain better performance than using smaller amounts of the same data for each individual context. We propose a data-driven methodology to find the best solution to this trade-off,

Without loss of generality, we can assume that a dialog state is represented by a vector of state variables. The possibly infinite states of a dialog system can be reduced by either clustering or ignoring certain variables; in fact, certain variables may have little or no impact on the input utterance distribution. For a given set of training data, one could experimentally, and automatically, evaluate optimal variable sets and, hence, training data subsets to generate context-adaptive grammars that maximize the overall performance. Initial results on data collected in the scope of several troubleshooting dialog systems indicate the applicability of this approach, although a full implementation of the concept is still not available, the main issue being the exponentially growing number of possible combinations of variable clusters per state.

References

W. Minker and S. Bennacef. 2004. *Speech and Human- Machine Dialog*. Springer, New York, USA.

D. Suendermann, J. Liscombe, K. Evanini, K. Dayanidhi, and R. Pieraccini. 2009. From Rule-Based to Statistical Grammars: Continuous Improvement of Large- Scale Spoken Dialog Systems. In *Proc. of the ICASSP*, Taipei, Taiwan.

Jonathan Bloom

jonathanb@speechcycle.com

Customization is a hard term to define, making this a hard topic to address. Instead of trying to provide a hard definition of customization, I will take a “prototype” approach to explain what it is. Then I will provide examples of how SpeechCycle employs customization.

Customization: What is it?

As far as I can tell, each instance of customization varies on three dimensions, with each dimension having a value that is considered more *representative* of what most people consider “customization”:

1. Time – Occurs across calls or during the course of an individual call. Across calls is considered more representative of customization.