Use and Acquisition of Semantic Language Model

Kuansan Wang **Ye-Yi Wang** **Alex Acero**
Speech Technology Group, Microsoft Research
One Microsoft Way, Redmond, WA 98052, USA
http://research.microsoft.com/srg

Abstract

Semantic language model is a technique that utilizes the semantic structure of an utterance to better rank the likelihood of words composing the sentence. When used in a conversational system, one can dynamically integrate the dialog state and domain semantics into the semantic language model to better guide the speech recognizer executing the decoding process. We describe one such application that employs semantic language model to cope with spontaneous speech in a robust manner. The semantic language model, though can be manually crafted without data, can benefit significantly from data driven machine learning techniques. An example based approach is also described here to demonstrate a viable approach.

1 Introduction

Any spoken language understanding system must deal with two critical issues: how to accurately infer user's intention from speech, and how to do it robustly amidst the prevalent spontaneous speech effects where users would inevitably stutter, hesitate, and self correct themselves on a regular basis. To address these issues, it has been proposed (Miller et al., 1994; Wang, 2000; Esteve *et al.,* 2003) that one can extend the statistical pattern recognition framework commonly used for automatic speech recognition (ASR) to the spoken language understanding (SLU) problem. The "pattern" to be recognized for ASR is a *string* of word, and for SLU, a *tree* of semantic objects that represent the domain entities and tasks that describe the user's intention. As is the case for ASR where a language model plays the pivotal role in guiding the recognizer to compose plausible string hypotheses, a pattern recognition based SLU relies on what is often called the semantic language model (SLM) to detect semantic objects and construct a parse tree from the user's utterance. Because the end outcome is a parse tree, SLM is usually realized using the structured language model techniques so that the semantic structure of the utterance can be included in modeling the language (Wang, 2000; Erdogan et al., 2002).

In this article, we describe an application of SLM in the semantic synchronous understanding (SSU) framework for multimodal conversational systems. A key idea of SSU is to immediately recognize and parse user's utterance, accepting only speech segments conforming to the prediction of SLM while the user is still speaking. Since the SLM can be updated in real-time during the course of interaction, irrelevant expressions, including the spontaneous speech, can be gracefully rejected based on what makes sense to the dialog context. In Sec. 2, we describe a study on the efficacy of SSU for a mobile personal information management (PIM) application called MiPad (Huang *et al.,* 2000). The SLM used there was manually derived with combined CFG and N-gram (Microsoft, 1999; Wang, 2002) by consulting the structure of the PIM back end without any user data. Obviously, the linguistic coverage of the SLM can be further improved with modern data-driven learning techniques. In Sec. 3, we describe one such learning technique that can utilize the manually crafted model as a bootstrapping template to enrich the SLM when suitable amount of training data become available.

2 SSU MiPad*

MiPad is a Web based PIM application that facilitates multimodal access to personal email, calendar, and contact information. MiPad users can combine speech commands with pen gestures to query PIM database, compose or modify email messages or appointments. We recently implemented a version of MiPad in HTML and SALT, taking the native support of SSU in SALT

* A video demonstration of SSU MiPad is available for download at http://research.microsoft.com/srg/videos/MiPadDemo_2Mbit.wmv

(Wang, 2002). Whenever a semantic object is detected, the PIM logic based on the current semantic parse is executed and the screen updated accordingly. The nature of SSU insures that the user receives immediate feedback on the process of SLU, and therefore can rephrase rejected and correct misrecognized speech segments. Studies (Wang, 2003) that contrast SSU with conventional turn taking based system show that, because SSU copes with spontaneous speech better, it elicits longer user utterances and hence fewer sentences are needed to complete a task. The highly interactive nature of SSU lends itself to more effective dynamic visual prompting, leading lower out of domain utterances. SSU also simplifies the confirmation strategy as every semantic object can be implicitly confirmed. Users have no trouble dealing with this strategy. In fact, users naturally correct and rephrase based on the immediate feedback, making their speech even more spontaneous. All these results are statistically significant. Finally and most intriguingly, users feel they accomplish tasks faster in the SSU system even though the through puts from both systems are statistically tied.

3 SLM Learning

SLU utilizes SLM to infer user's intention from speech. Before sufficient data make it practical to use machine learning techniques, SLM often has to be developed manually. The manual development process is labor-intensive, requires expertise in linguistics and speech understanding, and often lacks good coverage because it is hard for a developer to anticipate all possible language constructions that different users may choose to express their minds. The manually developed model is therefore *not robust* to extra-grammaticality commonly found in spontaneous speech. An approach to address this problem is to employ a robust parser to loosen the constraints specified in the SLM, which sometimes results in unpredictable system behavior (Wang, 2001). The robust parser approach also mandates a separate understanding pass from speech recognition. The results tend to be suboptimal since the first pass, optimizing ASR word accuracy, does not necessarily lead to a higher overall SLU accuracy (Wang and Acero, 2003b).

We have developed example-based grammar leaning algorithms to acquire SLM for speech understanding. It is shown (Wang and Acero, 2002) that a grammar learning algorithm may result in a semantic context free grammar that has better coverage than manually authored grammar. It is demonstrated (Wang and Acero, 2003a) that a statistical model can also be obtained by the learning algorithm, and the model itself is robust to extra-grammaticality in spontaneous speech. Therefore, a robust parser is no longer necessary. Most importantly, such a statistical SLM can be incorporated directly into the search algorithm for ASR, making a single pass, joint speech recognition and understanding process such as SSU possible. Because of that, the model can be trained directly to optimize the understanding accuracy. It is shown (Wang and Acero, 2003b) that the single pass approach achieved a 17% understanding accuracy improvement even though there is a signification word error rate increase, suggesting that optimizing ASR and SLU accuracy may indeed be two very different businesses after all.

References

Erdogan H. Sarikaya R. Gao Y. Picheny M. 2002. Semantic structured language models. *Proc. ICSLP-2002,* Denver, CO.

Esteve Y. Raymond C. Bechet F. De Mori R. 2003. Conceptual decoding for spoken dialog systems. *Proc. EuroSpeech-2003*, Geneva, Switzerland.

Huang X. *et al*. MiPad: A next generation PDA prototype. *Proc. ICSLP-2000*, Beijing, China.

Microsoft Corporation. 1999. Speech Application Program Interface (SAPI), Version 5.

Miller S. Bobrow R. Ingria R. and Schwartz R. 1994. Hidden understanding models of natural language. *Proc. 32nd Annual Meeting of ACL,* Las Cruces, NM.

Wang K. 2000. A plan-based dialog system with probabilistic inferences. *Proc. ICSLP-2000,* Beijing, China.

Wang K. 2002. SALT: A spoken language understanding interface for Web-based multimodal dialog systems. *Proc. ICSLP-2002,* Denver, CO.

Wang K. 2003. Semantic synchronous understanding for robust spoken language applications. *Proc. ASRU-2003*, St. Thomas, Virgin Islands.

Wang Y. 2001. Robust language understanding in MiPad. *Proc. EuroSpeech-2001*. Aalborg, Demark.

Wang Y. Acero A. 2002. Evaluation of spoken language grammar learning in the ATIS domain. *Proc. ICASSP-2002*, Orlando, FL.

Wang Y. Acero A. 2003a. Combination of CFG and N-gram modeling in semantic grammar learning. *Proc. EuroSpeech-2003.* Geneva, Switzerland.

Wang Y. Acero A. 2003b. Is word error rate a good indicator for spoken language understanding accuracy? *Proc. ASRU-2003,* St. Thomas, Virgin Islands.