A System for Multimodal Language Acquision

Sorin Dusan and James L. Flanagan

Center for Advanced Information Processing
Rutgers University
96 Frelinghuysen Road
Piscataway, NJ 08854, U.S.A.
{sdusan, jlf}@caip.rutgers.edu

Abstract. One of our goals in developing natural human-computer interfaces is to allow the computer to expand its vocabulary by learning new linguistic knowledge. This paper presents a system capable of acquiring new language by learning new words, phrases, sentences, and their semantics from users. The acquisition of the new linguistic knowledge at surface and semantic levels is done using multiple input modalities, including speaking, typing, pointing, touching or image capturing. The language knowledge is stored in a rule grammar and a semantic database. Both can be updated periodically with newly acquired language.

1 Introduction

Spoken dialogue systems require the implementation of speech technologies that include automatic speech recognition (ASR), text-to-speech (TTS), speech understanding and dialogue management. Dialogue systems based on unconstrained vocabulary of course offer the most natural interaction, but they are more difficult to implement than those based on constrained vocabulary stored in a rule grammar. A disadvantage of using dialogue systems based on rule grammars is that the developer cannot pre-program the rule grammar to account for all language preferences of users. Users find dialogue systems easier and more natural if they can change or adapt the allowed vocabulary and grammar according to their preferences.

In command and control applications based on speech, computers need to know the corresponding meanings of the understood words in order to perform appropriate actions. To enable computers to expand their vocabulary by learning new language, computers need to acquire linguistic knowledge at two levels: the *surface* level, represented by the syntax of the linguistic units; and the *deep* level, represented by the meaning of the linguistic units. For example, if a user wants to teach the computer the colour *burgundy*, this word and its semantics have to be acquired together by the computer. The semantics of this word could be the RGB attributes necessary for displaying this colour on the screen.

In a series of studies of language acquisition based on connectionist methods, [1], new words or phrases are acquired at the surface level and their corresponding meanings are determined by probabilistic associations with pre-programmed semantic actions.

Another study focused on the acquisition of linguistic units and their primitive semantics from raw sensory data, [2]. In that study the system learned new language by making associations between speech sounds representing words and their semantic representation acquired from a video camera.

A similar study, [3], focused on discovering useful linguistic-semantic structures from raw sensory data. The goal was to enable a robot to discover associations between words and different semantic representations obtained from a video camera.

A method of acquiring new linguistic units and their semantics using multiple input modalities was introduced in [4].

In this paper we present a computer system capable of acquiring language knowledge from multiple input modalities by learning new words, phrases, sentences and their semantic representations. The language knowledge is acquired from user in a supervised learning. The spoken dialogue system is suitable for command and control applications in which the vocabulary of the dialogue can be expanded and personalised by users according to their preferences.

2 Supervised Language Acquisition

In a spoken dialogue interface, the goal of language acquisition is to make the computer capable to understand new words and sentences and this can be accomplished by learning new linguistic knowledge. The language acquisition method presented in this paper is based on supervised language learning. We adopted a constrained-grammar dialogue method because this is suitable for command and control applications on a computer.

Our language acquisition system is supported by a computer with a multimodal interface based on a microphone and speakers for speech, a keyboard for typing, a mouse for pointing, a pen tablet for drawing and touching, a CCD camera for image capturing and a display for graphics and text. The concept of storing the language knowledge in two separate blocks is presented in Fig. 1.

The system interprets users' utterances according to allowed sentence structures stored in the rule grammar, and executes different actions according to information stored in the semantic database. The system contains a speech recognition engine and a text-to-speech engine. The rule grammar and the semantic database are stored in two different files on a hard disk from which they are loaded into computer memory. When the system detects unknown words and the user provides the corresponding semantic representation, the system dynamically updates the rule grammar and the semantic database with the new linguistic knowledge. At the end of the application, the user has the option to save permanently the updated rule grammar and semantic database in corresponding files on the hard disk.

The language acquisition takes place in real time during user-computer interaction. The computer identifies new words in the user's utterances and asks the user for the semantics of these words. The user can provide the semantics through multiple input modalities. In addition to adapting the vocabulary by

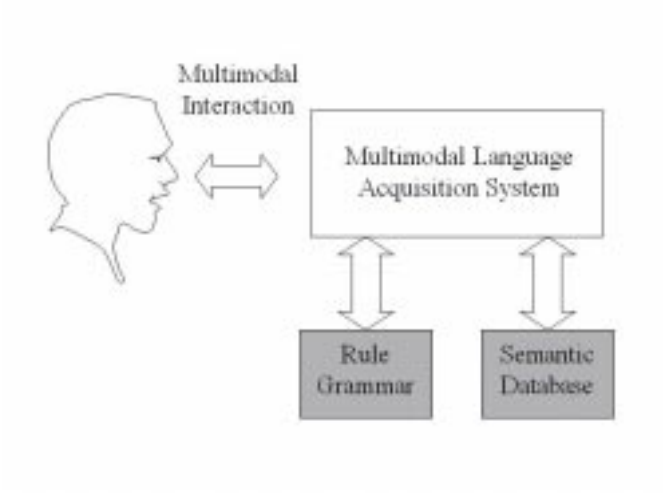

Fig. 1. Storing the language knowledge

learning new linguistic units and their semantics, the user can personalize the system's vocabulary by using synonyms or different names for already known terms.

3 Computer Representation of Language Knowledge

In command and control applications based on speech, computers need to associate to words specific meanings in order to execute appropriate actions. In our system the acquisition of language knowledge is done at two different levels: surface level and deep level. At the surface level, language is represented by knowledge of the vocabulary, syntax and grammar. We store this knowledge in a rule grammar. At the deep level, language is represented by knowledge of the meanings of linguistic units. We store this knowledge in a semantic database. The rule grammar and the semantic database are stored in two separate files on a hard disk from where they are loaded into the computer memory when the application starts. During the acquisition of new language knowledge, the rule grammar and the semantic database are updated in the computer memory. At the end of the application the user has the option to save permanently the acquired language in the corresponding files on hard disk.

3.1 Rule Grammar

A simple way to store the surface-level language knowledge into computers is in a rule grammar file. A rule grammar defines the allowed sentence structures by a set of production rules. We specify allowed sentence structures in a rule grammar, organised as a semantic grammar, [5]. In this format the nonterminal symbols represent semantic classes of concepts, such as *colours*, *fruits* and *geometric shapes*, and the terminal symbols represent concept words such as *yellow*, *apple* and *rectangle*.

The rule grammar can be dynamically updated by adding new words or phrases in semantic classes or by adding new production rules. A linguistic unit integrated into a semantic class has a corresponding semantic object stored in the semantic database.

3.2 Semantic Database

Semantic interpretation of utterances for performing necessary actions is based on knowledge stored in the semantic database. This database contains a set of semantic objects that describe the meaning (or a semantic representation) for each concept stored in each class in the rule grammar. This semantic database can be dynamically updated with new objects consisting of semantic representations of new linguistic units.

The semantic objects are created using the concept of object-oriented programming and they are instances of classes. The semantic representation stored in such an object defines the computer knowledge and interpretation of the corresponding linguistic unit. For example, the semantic object corresponding to the word *blue* included in the semantic class *colours*, has semantic representation defined by the RGB attributes (0, 0, 255). These attributes represent all computer knowledge regarding the meaning of this colour. Another example is the semantic object for the word *square*. In this case the object contains a pointer to a regular polygon and an attribute equal to 4 representing the number of sides. All characteristics of a *regular polygon* are thus inherited by the *square* semantic object. The semantic representation necessary to build these objects is either pre-programmed or acquired from the user through multiple input modalities.

4 Multimodal Language Acquisition

The main way for acquiring language in our system consists in teaching the computer new words and the corresponding semantics using multiple input modalities. Another way is to acquire new sentences or language rules through typing. A detailed block diagram of the system is given in Fig. 2.

The user's utterances are converted to text strings by an automatic speech recognition (ASR) engine and then analysed by a Language Understanding module, containing a Parser, a Command Processor, a Rule Grammar and a Semantic Database. Initially the ASR uses a language model derived from the Rule Grammar. The allowed utterances are converted to text at ASR output 1. Then they

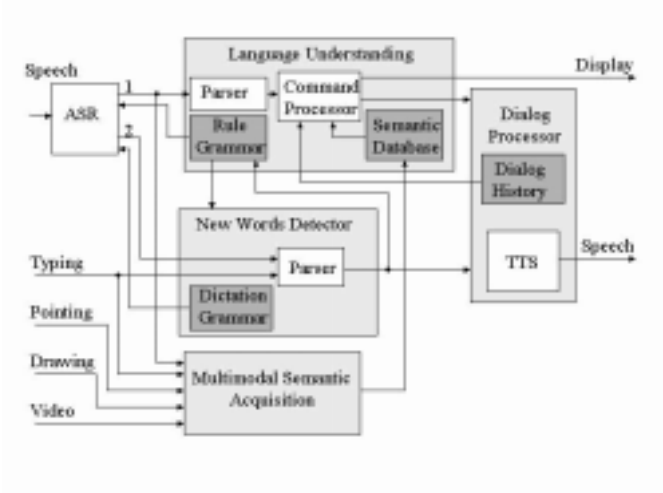

Fig. 2. Multimodal language acquisition

are parsed and executed by the Command Processor or forwarded to the Dialog Processor to provide appropriate answers through synthetic voice. The Command Processor also displays the results on the screen. The Dialogue Processor module includes a Text-To-Speech synthesizer and a Dialogue History necessary to solve ambiguities from the context of the dialogue. When the user's utterances contain unknown words or phrases, the ASR engine does not provide any text at its output 1. In this case the ASR switches the language model to one derived from the Dictation Grammar. Then these utterances are converted to text strings at ASR output 2 and are applied to the New Words Detector. This module contains a Dictation Grammar and a Parser. The Dictation Grammar contains a very large vocabulary of words and allows the ASR to recognise more unconstrained utterances. The Parser analyses these utterances according to all allowed words stored in the Rule Grammar and detects in these utterances unknown words or phrases. Upon the detection of new linguistic units, this module issues a signal to the Dialogue Processor that asks the user by synthetic voice to provide a semantic representation of the new words or phrases.

The user can provide a semantic representation using multiple modalities. For each new linguistic unit the semantic representation is captured by the Multimodal Semantic Acquisition module that creates a new semantic object. After the user has provided semantics for the new words or phrases, the new linguistic units are stored in the rule grammar and the corresponding semantic objects are stored in the semantic database.

Another means to acquire language is by teaching the computer by typing a whole new sentence and the corresponding computer action. The new sentence is stored in the rule grammar and the computer action must be based on a combination of known actions. For example, one can type in the new sentence 'Double the radius' and the corresponding computer action 'radius multiplication 2', where the word 'double' is unknown, but the words 'the', 'radius', 'multiplication' and '2' are known.

5 Experiments

Our multimodal language acquisition system was implemented on a personal computer running the Windows 2000 operating system by creating a graphic application with multiple input modalities including speaking, pointing with a mouse, typing on a keyboard, drawing on a pen tablet with a stylus and capturing images with a CCD camera. In this application, users can use spoken commands to display, move, delete and rotate graphical objects on the screen, or to assign or change the values of different application variables. The system is able to identify in the user's commands new words or phrases and to ask the user to provide the corresponding semantic representation. Upon receiving the semantics of the new linguistic units, the system stores the surface-level and semantic-level of information of the new linguistic knowledge and is able to recognize and understand the new terms in the future user's utterances. Users can also teach the computer system new sentences and sentence rules by typing.

This application is implemented in Java and the rule grammar is written in Java Speech Grammar Format (JSGF) specification. The automatic speech recognition and text-to-speech engines are from the IBM ViaVoice Release 8 Professional Edition. The dictation grammar includes a vocabulary of 160,000 words. The initial language, recognized and understood by the system, is preprogrammed in the rule grammar and the semantic database. The initial rule grammar consists of 30 sentence rules and 25 nonterminal symbols representing concept classes including display variables, actions, arithmetic operations, colours, 2D geometric shapes, regular polygons, graphic images, drawings, user names, etc. The initial semantic database contains semantic objects derived from concept classes, including colours such as red, green and blue, geometric shapes such as circle and line, regular polygons such as triangle and square, display variables such as radius, width and height, etc.

This application has been tested for different tasks. Two of these experiments are presented in the following. In a computer graphics task a user is able to teach the system many different colour names such as yellow, pink, mustard, burgundy, etc. and the corresponding computer representations. Also the computer is able to acquire from user new words describing regular polygons such as pentagon, hexagon or octagon, and a large number words associated with drawings, such as door, window, house, fence, cloud, tree, mountain, dog, cat, eye, nose, mouth, etc. The semantic representation of the new words can be structured on the already known concepts, for example a head can be described as a drawing containing

two eyes, a nose, a mouth, two ears, etc. The number of words the system can learn by associating them with graphic images or drawings is extremely large and it is practically limited only by the memory of the system. Users also taught the computer new language rules as described above.

In a different task, the user draws military icons or symbols for representing different words such as tank, helicopter, submarine, jeep, truck, barrack, ammunition, soldier, major, colonel, general, etc. and teaches the computer the corresponding names. Then using speech and pointing the user is able to place, move, rotate and delete these graphic symbols on a topographical map simulating a mission planning exercise. For example, the user can say 'Rotate this tank 45 degrees', 'Move this helicopter here', 'What is the position of this submarine?', and simultaneously point the cursor on the screen to the corresponding objects or positions using the mouse.

6 Conclusion

We present a multimodal language acquisition system capable of acquiring new language knowledge from users. The system can expand its vocabulary by adding new words or phrases in the rule grammar and by creating and storing into a semantic database new objects containing the corresponding semantic representation. An alternative method of acquiring new language knowledge is by typing new sentences and their corresponding computer actions. The vocabulary of the system can also be personalised by users by providing synonyms or different proper names to already-known terms.

This research was supported by the National Science Foundation under the Knowledge and Distributed Intelligence project, grant NSF IIS-98-72995.

References

1. Gorin, A: On automated language acquisition, Journal of the Acoustical Society of America, **97** (6), (1995) 3441–3461
2. Roy, D. K.: Learning Words from Sights and Sounds: A Computational Model, Ph.D. Thesis, MIT (1999)
3. Oates T.: Grounding Knowledge in Sensors: Unsupervised Learning for Language and Planning, Ph.D. Thesis, MIT (2001)
4. Dusan, S. Flanagan, J. L.: Human Language Acquisition by Computers, in Proceedings of the International Conference on Robotics, Distance Learning and Intelligent Communication Systems, WSES/IEEE, Malta (2001) 387-392
5. De Mori, R.: Recognizing and Using Knowledge Structures in Dialog Systems, In Proceedings of IEEE-ASRU99, Keystone CO (1999) 297-307