Verb Processing in Spoken Commands for Household Security and Appliances

Ioanna Malagardi[1] and Christina Alexandris[2]

[1] Educational & Language Technology Laboratory
Department of Informatics & Telecommunications
National and Kapodistrian University of Athens, HELLAS
imalagar@di.uoa.gr
[2] National and Kapodistrian University of Athens, HELLAS
calexandris@gs.uoa.gr

Abstract. The present paper concerns the handling of verbs in the Speech Recognition Module of an HCI system for the remote control of household security and the operation of household appliances. The basic language used is Modern Greek, but the system's design includes the basis of a multilingual extension for the use of the system by native-speakers of other languages. The human- computer communication must preferable to be accomplished in natural language. Some methods of Artificial Intelligence can contribute to the solving of the natural language processing problems. The target for a multilingual extension of the system has imposed the restrictions that commands are kept simple and referring expressions such as deictic noun phrases and pronouns as well as anaphoric expressions are avoided. The interaction with the system is strictly based on dialogs with restricted options in order to increase the feasibility of the speech interface.

Keywords: speech recognition, natural language processing, motion verbs, interlinguas.

1 Introduction

The present paper proposes the architecture of a system for the handling of verbs in the Speech Recog nition Module of an HCI system for the remote control of household security and the operation of household appliances. The basic language used is Modern Greek, but the system's design includes the basis of a multilingual extension for the use of the system by native-speakers of other languages, to fulfil the needs of the multinational workforce in Greece today.

The target for a multilingual extension of the system has imposed the following two restrictions: (1) Commands are kept simple and Referring Expressions such as deictic noun phrases (i.e "this window"), deictic pronouns (i.e. "this, that") [6] and pronouns related to anaphoric expressions ("it", "they") are avoided. (2) The interaction with the system is strictly based on dialogs with restricted options. Thus, dialog management does not involve processing conversations with the system [4].

C. Stephanidis (Ed.): Universal Access in HCI, Part II, HCII 2009, LNCS 5615, pp. 92–99, 2009.

Commands are restricted to simple orders, in the form of imperatives and expressing three types of actions. The first action is the movement of an object (change of position), the second action is the opening or closing of some objects (change of state) and the third action is to put one object on another object (change of relation). Even for the simple graphical representation in the computer's screen, we have to consider the physical attributes of the objects and principles of geometry and physics. Here, the actions concerned only involve actions related to change of state.

2 Understanding and Managing Verbs

Understanding the imperatives requires understanding the meaning of actions such as "open", "close", "put" and the meaning of prepositional words such as "on". One integrates the meanings of the constituents and produces a meaning of sentence as a whole, taking pragmatic factors into consideration where appropriate. Having done so, the system constructs a plan for execution of the task in the environment. Only then can the system perform the action in the given environment.

Among the many issues involved in the comprehension of imperatives in a physical domain and the execution of underlying tasks, it is of crucial importance to represent the meanings of verbs and prepositions in order to characterize underlying actions. Thus, we describe a simple representation in which movements denoted by action verbs can be expressed in a manner that can be implemented in terms of a computer program. Suppose an agent is asked to perform the following commands in a suitable environment: "Open the door", "Open the bottle", "Close the box", and "Put the book on the desk". Each of these sentences specifies an underlying task requested of an agent. In order to perform the task, the performing agent has to "understand" the command. Understanding the imperatives requires understanding the meaning of actions such as "open", "close", "put" and the meaning of prepositional words such as "on". The agent must integrate the meanings of the constituents and produce a meaning of the sentence as a whole, taking pragmatic factors into consideration where appropriate. Having done so, one must construct a plan for execution of the task in the given environment. Only then can the agent perform the action. All of the above steps need to be followed, regardless of whether the agent is human or program-controlled such as an animated agent in a computer graphics environment or a robotic agent [8].

2.1 The Complexity Factor in Expressing Motion

Motion can be indicated by a verb either directly or indirectly. The simplest way to specify motion of an object is by using a verb that specifies motion in a straightforward manner. An example verb is "move" as used in the sentence "Move the chair from the wall to the table". It simply directs the system to execute a motion with the chair as the affected object.

Indirect specification of motion can be achieved in two ways: in terms of geometric goals, or in terms of a force. Indirectly specifying motion in terms of a goal involving physical relationship among objects is quite common among verbs. Consider the sentence "Put the bottle on the table". The instruction requires that a physical object be moved (i.e., the bottle) with a goal to establish a physical relationship (the relationship of "on") between it and another physical object (i.e., the table). The performance of such an instruction demonstrates that the goal of establishing a physical relationship drives the motion of the first object.

For verbs such as “put” that specify motion in terms of a geometric goal, properties of the objects that participate in the underlying action are of crucial importance. Except these verbs, there is another way to specify motion indirectly without using these verbs. This is by specification of a force rather than the actual motion itself. In these cases too, we have to focus primarily on physical characteristics of the actions that underlie motion verbs. In order to do so, we need to obtain physically realizable representations for the meanings of such verbs.

One source of the multiplicity of meaning of a command is the multiplicity of the senses of a word as recorded in a dictionary. Another source is the possibility of an object to be placed on a surface in different ways. For instance, when the user submits a command, the agent, in order to satisfy the constraints of the verb, he may ask for new information and knowledge about objects and verbs, which may be used in the future. In this case, a machine-readable dictionary would be used, providing the definition of the verbs [8].

For example, if the user enters the command “open the door”. The agent isolates the words of the command and recognizes the verb “open” and the noun phrase “the door”. The verb “open” appears in the lexicon with a number of different definitions. For example, in the LDOCE [12] we find, among others, the senses of “open” a: to cause to become open, b: to make a passage by removing the things that are blocking it. The agent finds in the knowledge base that there are two alternative ways of interpreting the verb “open”, using either a “push” or a “pull” basic motion. Then, it selects the first one and asks the user if this is the right one. When the user enters a “Yes” answer, this is recorded in the knowledge base and the process terminates. When the user enters a “No” answer, the process continues trying sequentially all the available sides of the book until a “Yes” answer is given by the user [9].

The above-presented scenario involving a knowledge base and a sequence of questions and answers performed between the user and the agent (system) may provide a more rigorous and sublanguage-independent approach in respect to motion verbs and their arguments (objects), however, it entails difficulties in its implementation in spoken and multilingual applications.

3 Input Management

In the present system, physically realizable representations for the meanings of motion-verbs concern actions to be performed in respect to household security and appliances. The requested actions comprise user-queries or system-output concerning (1) the performance of an action [7], for example “Lock all the windows”/ “All windows are locked”, or a (2) check [7], for example “Is the central-heating on?”/ “The central-heating is turned off”.

The set of lexical entries in state and action types may be paired with phrases or expressions initiating sentences constituting queries in respect to (a) actions (“Action”) that the user asks to be performed or (b) queries in respect to objects that the user wishes to be checked (“Check”) respectively.

Table 1. Relation of Use Case, Function and Code

Use Case	Function	Code
Use Case 1:	House-Security Function	HOUSE
Use Case 2:	Appliances-Control	APPLIANCES

Input management for the Speech Recognition Module is based on the use of keyword lists. Keyword lists are linked to user input control, in the form of keyword-groups. Keyword recognition includes a number of yes-no question sequences of a Directed Dialog [15], [16]. The use of directed dialogs and yes-no questions aims to the highest possible recognition rate of a very broad and varied user group. Additionally, the use of selected keywords allows the efficient handling of ambiguous "Multitasking" verbs, typically occurring in Greek [3].

"Multitasking" verbs are related to multiple semantic meanings and used in a variety of expressions, existing in Modern Greek, and possibly in other languages as well. For example, in Greek, at least in the sublanguage related to the communication context of commercial activities, the semantically related verbs "buy", "get" and "purchase" may be used in similar expressions as the (primitive) verbs "is" and "have" as well as the verbs "give" and "receive" to convey the same semantic meaning from the speaker [14], [3].

Keywords constituting the actual elements recognized by the system may divided into three main categories: (a) Elements consisting keywords that are mapped in respect to closed and relatively small lists, (b) Elements that are mapped in respect to open databases, such as names, (addresses and locations may be added) (c) Elements consisting numbers that may include information such as quantity, address or date.

The type of input recognized by the Speech Recognition Module is attempted to focus on (at least) two types of keywords within the Speaker's utterance, namely the type of action requested by the Speaker and the type of object or activity related to the requested type of action. "Time" is an optional parameter added to this basic form. This approach may be formally described in its basic from as a Template related to the type of content of the utterance: [(OBJECT) + (ACTION-TYPE)].

Table 2. Keywords categories recognized by the Speech Recognition Module (Basic form)

Category: Object	Category: Action
[(OBJECT) + (ACTION-TYPE)]	[(OBJECT) + (ACTION-TYPE)
OBJECT:	ACTION-TYPE:
(HOUSE-FEATURES)	= OPEN, CLOSE, START, STOP,
(APPLIANCE-TYPE)	CHECK

The actual main components of the speakers response, constituting keyword categories, may be described, in the present application, as a sublanguage-specific set of categories (closed lists) (a) such as: (HOUSE-FEATURES), (APPLIANCE-TYPE), (ACTION-TYPE) and open-categories (b) for names, (possibly addresses and place names (PLACE), in a future extension of the system) as well as keyword categories are related to temporal information and quantitative expressions (Time).

Keywords grouped under ACTION-TYPE involve expressions related to activities such as activating the alarm or checking if the power supply is turned off. Specifically, keywords constituting ACTION-TYPE are expressions related to requested actions to be performed in respect to household features and appliances, for example "Turn on TV", "Open the door" or "Turn off the oven". Additionally, Keywords constituting ACTION-TYPE include expressions related to checking the operation of household features and appliances, for example "Is the gas switched off?" and "Is the alarm on?". The logical relations of the two types of keywords to be recognized by the Speech Recognition Module and the actual words related to each closed list are described by Table 2 and Table 3 respectively.

Table 3. Words related to the closed lists and open lists of Keyword categories recognized by the Speech Recognition Module

Keyword category	Function type	Keyword list
ACTION-TYPE	OPEN, CLOSE START STOP	OPEN = activate, activated, open, opened, running, switch-on, switched-on, turn-on, turned-on CLOSE = ,close, closed, de-activate, de-activated shut, shut-down, switch-off, switched-off, turn-off, turned-off turn-off, stop START = run, start, started, start, begin STOP = stop, stopped, pause, paused
ACTION-TYPE	CHECK	CHECK = inform, check, see, look
Objects (OBJECT):	HOUSE-FEATURES APPLIANCE-TYPE:	HOUSE-FEATURES: doors, windows, alarm, garage-door, door-lock, camera APPLIANCE-TYPE: central-heating, gas, electricity, power-supply, lights, oven, refrigerator, washing-machine, television

Time:	DAY-OF-WEEK
	RELATIVE-TIME
	CLOCK
	DATE
		DAY-OF-WEEK = Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday, Weekend
		RELATIVE-TIME = today, tomorrow, yesterday, CLOCK = twelve o'clock, half past two, ten fifteen
		DATE =February the eighteenth, March the third

Lexical entries composed of more than one word that have to be processed by the system as a singular expression are presented with a dash "-" between the components.

The limited set of lexical entries is chosen according to the criteria of simplicity, directness in order avoid as much as possible the occurrence of (1) ambiguities in respect to the speech recognition component and (2) complications in the user's/hearer's understanding of the system output constituting natural or synthetic speech [1],[2]. For example, expressions such as "activate" (a device, a program), although are in general practice regarded as highly appropriate and correct by professionals and the computer literate, may have the effect of rather unusual or even incomprehensive to a remarkable percentage of users like the elderly or non-native speakers.

Therefore, the present system also allows the recognition of simpler expressions such as “open” to be mapped to the same command or information as a more “appropriate” expression such as “activate”.

This basic set of lexical entries and respective dialogs allows the possibility of additional development according to the needs of the User Cases utterances and respective lexical additions to keyword groups.

4 User-Friendly System Output

System output in respect to the information on the objects is related to the limited set of lexical entries in state and action types described above. The above-described pairing of lexical entries with phase- or expression types also foresees and allows the default handling of less than perfect speaker’s utterances, since spoken language is characterized by fragmented syntactical structures. The chances of ungrammatical pairings of lexical entries with phase- or expression types are accounted for, however, are predicted to be very limited with native speakers.

The effort must be made for the utterances produced by the system to be (1) clear and unambiguous towards the user but at the same time to be (2) friendly and natural-sounding and, in addition, to contain expressions that, from a semantic aspect, (3) constrain the range of the user’s possible responses to a minimum, thus restricting as far as possible the probability of ambiguities and misinterpretations regarding user input.

Thus, the system must be compatible to the criteria of successful operation at the Utterance Level (Informativeness, Intelligibility, Metacommunication handling i.e. repetition, confirmation of user input/pauses), the Functional Level (Ease of use/functional limits, Initiative and interaction control, processing speed/smoothness) and the Satisfaction Level (Perceived task success, comparability of human partner, trustworthiness) [13].

The Speech-Act oriented approach in the steps of the dialog structure for spoken technical texts are targeted to meet the requirements of “Precision”, “Directness” and “User-friendliness”, summarizing the criteria of Informativeness, Intelligibility and Metacommunication handling on the Utterance Level (Question-Answer-Level) [13], the Functional Level (initiative and interaction control) and the Satisfaction Level (perceived task success, comparability of human partner and trustworthiness) [13].

5 Multilingual Extension of the System

Although keyword-group user-input may vary according to the language or even in respect to the user, this type of input cannot deviate considerably from being restricted and hence, manageable for multilingual applications and allowing minimum interference of language-specific factors. In an attempt to meet the needs of the diverse community of foreign residents, the present system allows the use of Interlinguas (ILTS), to be used as semantic templates for a possible multilingual extension of the present dialog system.

The Interlinguas are designed to function within a very restricted sublanguage, with a rigid and controlled dialog structure based on Directed Dialogs, most of which involve Yes-No Questions or questions directed towards Keyword answers. The structure of the proposed Interlinguas is based on a strategy for filtering user-input for the efficient handling of both ambiguous and "multi-use" expressions used for expressing multiple types of information [3].

Traditional ILTS [5],[10],[11] are constructed around a verb-predicate signalizing the basic semantic content of the utterance, the so-called "frames". Thus, for example, the utterance "I am booked for Friday" is signalized by the frame "booked". In the present application, the role of the "frame" in the Interlingua structure is weakened and the core of the semantic content is shifted to the lower level of the lexical entries (Table 4). The "frame" level will not signalize the meaning of the sentence: This task will be performed by the lexical entries. The proposed Basic Interlinguas [3] may be characterized to be more of Interlinguas with an accepting or rejecting input function [3] rather than the traditional Interlinguas with the function of summarizing the semantic content of a spoken utterance.

Table 4. Basic Interlingua

"Frame" type	Keyword categories	Object type
ACTION	WHAT (OBJECT) WHERE (PLACE) WHEN (TIME)	OBJECT (HOUSE)
CHECK	WHAT (OBJECT) WHERE (PLACE) WHEN (TIME)	OBJECT (APPLIANCE)

6 Conclusions and Further Research

The processing of spoken commands involving movement by an HCI system intended for the broad public and with an envisioned extension to multilingual applications entails a well-structured approach in the Design Phase. For the remote control of household security and operation of household appliances, the above presented approach, facilitates Speech Recognition, thus, contributing to the quality, usability and safety of the system. The next step is the integration of the above-proposed strategy in the Speech Recognition Module, its evaluation and subsequent adaptation to multilingual applications.

References

1. Alexandris, C.: Word Category and Prosodic Emphasis in Dialog Modules of Speech Technology Applications. In: Botinis, A. (ed.) Proceedings of the 2nd ISCA Workshop on Experimental Linguistics, ExLing 2008, Athens, Greece, August 2008, pp. 5–8 (2008)
2. Alexandris, C.: Show and Tell: Using Semantically Processable Prosodic Markers for Spatial Expressions in an HCI System for Consumer Complaints. In: Jacko, J.A. (ed.) HCI 2007. LNCS, vol. 4552, pp. 13–22. Springer, Heidelberg (2007)

3. Alexandris, C.: The CitizenShield Dialog System in Multlingual Applications. In: Proceedings of the National Conference in Knowledge Management and Governing Systems, Hellenic Society of Systemic Studies – HSSS 2007, Pireus, Greece, May 12-14 (2007)
4. Bos, J., Ota, T.: A spoken language interface with a mobile robot. Artificial Life and Robotics 11, 42–77 (2007)
5. Dorr, B., Hovy, E., Levin, L.: Machine Translation: Interlingual Methods. In: Brown, K. (ed.) Encyclopedia of Language and Linguistics, 2nd edn., ms. 939 (2004)
6. Foster, M.E., Gurman-Bard, E., Guhe, M., Hill, R.L., Oberlander, J., Knoll, A.: The Roles of Haptic-Ostensive Referring Expressions in Cooperative, Task-based Human-Robot Dialogue. In: Proceedings of HRI 2008, Amsterdam, The Netherlands, March 12-15, 2005 (2008)
7. Heeman, R., Byron, D., Allen, J.F.: Identifying Discourse Markers in Spoken Dialog. In: Proceedings of the AAAI Spring Symposium on Applying Machine Learning to Discourse Processing, Stanford (March 1998)
8. Kontos, J., Malagardi, I., Trikkalidis, D.: Natural Language Interface to an Agent. In: EURISCON 1998 Third European Robotics, Intelligent Systems & Control Conference Athens. Published in Conference Proceedings "Advances in Intelligent Systems: Concepts, Tools and Applications", pp. 211–218. Kluwer, Dordrecht (1998)
9. Kontos, J., Malagardi, I., Bouligaraki, M.: A Virtual Robotic Agent that learns Natural Language Commands. In: 5th European Systems Science Congress. Heraklion Crete. Res. Systemica., vol. 2 (October 2002) (special issue), `http://www.afscet.asso.fr/resSystemica/accueil.html`
10. Levin, L., Gates, D., Lavie, A., Pianesi, F., Wallace, D., Watanabe, T., Woszczyna, M.: Evaluation of a Practical Interlingua for Task-Oriented Dialogue. In: Proceedings of ANLP/NAACL-2000 Workshop on Applied Interlinguas, Seattle, WA (April 2000)
11. Levin, L., Gates, D., Wallace, D., Peterson, K., Lavie, A., Pianesi, F., Pianta, E., Cattoni, R., Mana, N.: Balancing Expressiveness and Simplicity in an Interlingua for Task based Dialogue. In: Proceedings of ACL 2002 Workshop on Speech-to-speech Translation: Algorithms and Systems, Philadelphia, PA (July 2002)
12. Longman Dictionary of Contemporary English, The up-to-date learning dictionary. Editor-in-Chief Paul Procter. Longman group Ltd., UK (1978)
13. Moeller, S.: Quality of Telephone-Based Spoken Dialogue Systems. Springer, New York (2005)
14. Nottas, M., Alexandris, C., Tsopanoglou, A., Bakamidis, S.: A Hybrid Approach to Dialog Input in the CitzenShield Dialog System for Consumer Complaints. In: Proceedings of HCI 2007, Beijing China (2007)
15. Williams, J.D., Witt, S.M.: A Comparison of Dialog Strategies for Call Routing. International Journal of Speech Technology 7(1), 9–24 (2004)
16. Williams, J.D., Poupart, P., Young, S.: Partially Observable Markov Decision Processes with Continuous Observations for Dialogue Management. In: Proceedings of the 6th SigDial Workshop on Discourse and Dialogue, Lisbon (September 2005)