An Hybrid Approach for Spoken Natural Language Understanding Applied to a Mobile Intelligent Robot

Mário Rodrigues[1], António Teixeira[2], and Luís Seabra Lopes[2]

[1] Escola Superior de Tecnologia e Gestão de Águeda, Universidade de Aveiro, 3754-909 Águeda, Portugal
[2] Departamento de Electrónica e Telecomunicações/IEETA, Universidade de Aveiro, 3810-193 Aveiro, Portugal

Abstract. The greater sophistication and complexity of machines increases the necessity to equip them with human friendly interfaces. As we know, voice is the main support for human-human communication, so it is desirable to interact with machines, namely robots, using voice. In this paper we present the recent evolution of the Natural Language Understanding capabilities of Carl, our mobile intelligent robot capable of interacting with humans using spoken natural language. The new design is based on a hybrid approach, combining a robust parser with Memory Based Learning. This hybrid architecture is capable of performing deep analysis if the sentence is (almost) completely accepted by the grammar, and capable of performing a shallow analysis if the sentence has severe errors.

1 Introduction

Recent developments For these robots to emerge it's essential the development of natural language interfaces, regarded as the only acceptable for a high level of interaction [1]. Voice allows hands free communication even without visual contact, great advantages if the machine is a mobile robot. In this line of research, we are developing a mobile intelligent robot named Carl. Currently the development of such robots is still a challenge due to limitations of current technologies and the nature of input information: speaker independent speech recognition is not very reliable, even in quiet environments; performance degrades considerably with background noise; spontaneous spoken language is often highly disfluent.

The use of spoken language interfaces in robots requires analysis components robust to various types of disfluencies that can extract the most complete interpretation possible from a given input, grammatically correct or not [2]. It would be of little application a robot that in all the cases the speech recognizer makes an error doesn't react.

In this paper we present the current status of the spoken natural language interface of robot Carl focusing in our implementation of a robust Natural Language Understanding (NLU) module. We start by a brief presentation of the robot, in Section 2. Next, in Section 3 we describe the previous NLU modules, for an easier understanding of our new approach. The new approach is presented in Section 4, and results obtained with these new developments are the object of Section 5. Paper ends with results discussion and indications of ongoing and future work.

Rodrigues M., Teixeira A. and Seabra Lopes L. (2004).
An Hybrid Approach for Spoken Natural Language Understanding Applied to a Mobile Intelligent Robot.
In *Proceedings of the 1st International Workshop on Natural Language Understanding and Cognitive Science*, pages 145-150
DOI: 10.5220/0002667001450150

2 Robot Carl

Carl is based on a Pioneer 2-DX indoor platform from ActivMedia Robotics. This mobile platform includes an on-board Pentium based computer. A fiberglass structure was added on the top of the mobile platform, making Carl 1.10 m tall. It carries a laptop computer, a microphone array, a speaker and a webcam.

With this platform we are developing an autonomous robot capable of navigate in unstructured environments, making decisions, executing tasks and learning. Human-robot communication is achieved through spoken and written language dialog as well as touch interactions. High-level reasoning, including inductive and deductive inference, is mostly based on the Prolog inference engine.

2.1 Speech Interface Goals

The goal Carl's speech interface is to map each sentence to one of the performatives from the Human Robot Communication Language (HRCL) (defined in [1]). For example, a sentence like "Turn left." should result in performative `Achieve()`.

Carl is a robot designed to wander autonomously in an indoor environment interacting with English speaking humans, so the speech interface must be robust enough to allow the robot to engage communication with any person anywhere.

The communication with any person implies a user independent speech recognizer. The voice models of such recognizers are not so specific as the ones of user dependent systems and their performance is not as good, but they are more suitable to be used by a broader population. Communicating anywhere implies that Carl has to operate in uncontrolled environments like exhibition pavilions where the background noise level can be very high. This fact makes the speech recognition an even more difficult task to accomplish.

These conditions make the task of finding the appropriate HRCL performative from each spoken sentence a challenging problem. The first step is to determine the recognized sentence performative type. After that, semantic information must be extracted in order to fill the HRCL message to pass to the reasoning module of the robot [3].

3 Previous Approaches

In one of the first versions of the voice interface, IBM ViaVoice ASR performed speech recognition and the NLU module was built with CPK-NLP tools. Both speech recognition and NLU used a unification grammar defined in Augmented Phrase Structure Grammar (APSG) formalism.

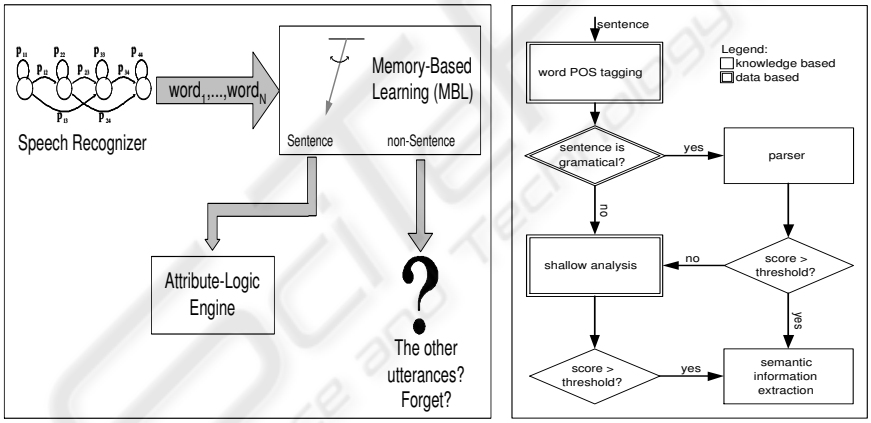
With this approach each recognized sentence had a valid analysis. However, codifying the linguistic knowledge as a set of rules is a labor-intensive task. Also, with the grow of the grammar it becomes increasingly difficult to maintain and expand the rule set. Another problem is that the speech recognition is performed based on a rigid set of possible sentences. If the user does not pronounce a sentence accepted by the grammar, it is not possible to recognize it. Obviously, the goal of developing a friendly easy to use robot was not accomplished, motivating a first change of approach.

The first step was to allow the recognizer to accept any word sequence. The approach with better tradeoff between flexibility and accuracy is to use an n-gram language model in the speech recognizer. The use of n-grams allows any sequence of words from the lexicon, while making some sequences more likely than others.

A new speech recognizer was selected, Nuance 8.0, and the grammar was built with bigrams trained from the complete set of utterances of the previous grammar. A new module based on the Attribute Logic Engine (ALE) [4]was developed. ALE still uses a grammar in its syntactic and semantic analysis.

4 New Approach

Having a system that is good at understanding correct sentences, we decided to develop a new system that makes use of the existing one in the cases it is adequate and uses a new sub-module to handle the deficiencies. To be able to apply such a divide to conquer approach, we first developed a module capable of deciding if the speech recognizer output was adequate to be processed by ALE. The system at this stage is represented in Fig. 1(a). After that, we worked on the semantic analysis of badly structured sentences. Then we replaced ALE by a more flexible parser.



(a) A first stage in the development (b) Complete version

Fig. 1. Our new natural language interface approach scheme

4.1 Sentence/Non-sentence Decision

For this sub-problem, a Memory-Based Learning (MBL) approach was adopted. The classification program TiMBL [5] was used. Words coming from the speech recognizer are complemented with part-of-speech (POS) tags. With 762 examples in the training set, correct decisions are above 90%. 3000 examples were enough to obtain 95% correct decisions [3].

This success motivated us to extend the importance and the amount of information extracted with MBL approach.

4.2 Robust Parsing

Another instance of TiMBL was used to extract information from the previously ignored sentences. However, MBL is not suitable for a deep syntactic analysis, and so, even for sentences that only slightly deviate from the grammar, it was only possible to perform shallow analysis. Also, traditional parsing algorithms as the ones of CPK-NLP and ALE are designed to parse only completely grammatical input sentences. It is desirable to have a tool that in the presence of a sentence with a small deviation from the grammar is able to perform a deep analysis: a robust parser.

Because the robust parsers are able to parse also completely correct sentences, such a tool could replace ALE. Tests are being made to a parser called LCFLEX[6]. LCFLEX is a flexible left-corner parser designed for efficient robust interpretation that uses the Lexical Functional Grammar (LFG) formalism for grammar definition. It supports word skipping, insertion of non-terminal grammar rules and selective flexible feature unification.

4.3 The New System

To build the syntactic structure of a sentence, the information used is the role of each word in the sentence and not the word itself so the lexical entries of the grammar used by LCFLEX are POS tags instead of words. The POS tags are assigned on-the-fly by two taggers: Brill Tagger [7] and LTPOS [8]. The reason for using two taggers is that no tagger has 100% accuracy. If both taggers assign the same tag, the system considers that the tag is correct; otherwise the system it treats specially the word and the sentence. Both taggers use the Penn Treebank tagset that has 36 tags, so the lexicon of our grammar has just 36 entries which allows us to have a very small, easy maintained grammar.

The new natural language interface architecture is represented in Fig. 1(b). First a module determines the POS tags of the words of the sentence. Then an instance of TiMBL decides if the sentence retrieved by the speech recognizer is grammatical or not. If both taggers agree in more than 75% of the tags and the sentence is grammatical, the sentence is passed to LCFLEX to perform a deep syntactic analysis, otherwise it is passed to another instance of TiMBL that makes a shallow analysis. Since the decision made by the first instance of TiMBL as an error of around 5% ([3]), LCFLEX can receive some incorrect sentences. Usually these sentences are almost grammatical and LCFLEX is capable of analyse them.

If the sentence syntactic structure returned by LCFLEX includes more then 75% of the total words of the sentence, the threshold, the analysis is passed to the semantics extraction module. Otherwise the system considers that the sentence has a great deviation from the grammar and the final analysis is shallow and made by the second instance of TiMBL. This instance of TiMBL can also ignore the sentence if it cannot get a valid analysis.

If the speech recognizer passes more than one hypothesis to the NLU module, in order to have the best global solution, every hypothesis is analyzed and the partial scores are collected. The partial scores are, if applicable, the percentage of equal tags assigned

by both taggers, first and second TiMBL instances scores and percentage of words included in sentence syntactic structure returned by LCFLEX. TiMBL scores are computed as the ratio between the number of neighbors of the most frequent valid sentence class and the number of neighbors of the most frequent class. In the end of the analysis, the partial scores are weighted with the speech recognition score and the best overall scored analysis is chosen.

With this hybrid architecture we have developed an interface capable of performing deep analysis if the sentence is completely or almost completely accepted by the grammar, and capable of performing a shallow analysis if the sentence has severe errors.

5 Some Results

In Table 1 we present a summary of a preliminary test of the new proposed approach. A total of 162 sentences from a randomly selected list generated from the ALE grammar was read aloud by one speaker in an environment with some background noise (a common research laboratory). In average each sentence had 5.53 words. The speech recognizer used was the Nuance with bigrams language model. The first two best alternatives were kept from Nuance. Only 13 times in these alternatives the intended sentence was present without errors. Speech Recognizer errors are presented in Table 1(a). The Word Error Rate (WER) shows that speech recognizer results are full of errors.

Table 1. Test results

(a) Speech Recognizer errors

replaced	deleted	inserted	WER
1.21	0.56	0.69	44.42

(b) Robust parser results

adequate to ALE	processed by LCFlex	processed by TiMBL	sentences with analysis
21	50	20	70

(c) Detailed analysis of sentence type results

	total	LCFlex	TiMBL
correct	52	37	15
incorrect	18	13	5

A little more than the 13 correct sentences at the speech recognizer output were adequate to be processed by ALE. The 8 ($21 - 13 = 8$) incorrect sentences processed by ALE came from subtle changes, especially in articles. At the end, Table 1(b), our system obtained analysis for 70, an 233.33% increase.

Regarding type of sentence, 52 of the 70 were correct. Partial results from TiMBL and LCFLEX are presented in Table 1(c). It is also interesting to look at the constructed semantic relations. In 209 constructed relations 156 were judged by a human as correct. Relations constructed for sentences accepted by ALE were correct in 59 of a total of 60 cases. The failure was also made by ALE due to the loss of a coordinative conjunction.

After passing all the processing, the system was incapable of extracting any information from 27 sequences.

6 Conclusion

In this paper we presented the recent evolution of the NLU module of the mobile intelligent robot Carl. This new module is based on a hybrid approach combining a robust parser with MBL-based modules to complement and select cases appropriate for the robust parser use. With this hybrid architecture we have developed an interface capable of performing deep analysis if the sentence is completely or almost completely accepted by the grammar, and capable of performing a shallow analysis if the sentence has severe errors.

Results of a preliminary test performed indicate that our new approach is capable of handling an increased number of word sequences coming from the speech recognizer. For the sequences grammatically correct or almost correct, the new system has a performance equivalent to the previous ALE based approach. Sentences with errors are handled mainly by LCFLEX, but MBL contribution is not irrelevant. Results from both ALE and MBL are around 74% correct regarding the type of performative, and aprox. 75% regarding constructed semantic relations. It is clear from the results that both components, MBL and the robust parser, perform a part of the task, contributing to the overall performance.

As an ongoing work, many evolutions are possible. We consider as prioritary the fine tuning of all the decision levels in the processing, improvement of the training material for MBL bases tasks, better control of the LCFLEX flexibility parameters.

References

1. L. S. Lopes and A. Teixeira. Teaching Behavioral and Task Knowledge to Robots through Spoken Dialogues. In S. LuperFoy and D. Miller, editors, *My Dinner with R2D2: Natural Dialogues with Practical Robotic Devices*, AAAI-2000 Spring Symposium Series, pages 35–43. Stanford University, Stanford, California, March 2000.
2. C. P. Rosé and A. Lavie. Balancing Robustness and Efficiency in Unification-Augmented Context-Free Parsers for Large Pratical Applications. In J.-C. Junqua and G. van Noord, editors, *Robustness in Language and Speech Technology*. Kluwer Academic Publishers, 2001.
3. L. Seabra Lopes, A. Teixeira, M. Rodrigues, D. Gomes, C. Teixeira, L. Ferreira, P. Soares, J. Girão, and N. Sénica. Towards a Personal Robot with Language Interface. In *Proc. EuroSpeech*, pages 2205–2208, 2003.
4. B. Carpenter and G. Penn. ALE - The Attribute Logic Engine User's Guide. Technical report, Department of Computer science of University of Toronto, 10 King's Colledge Rd., Toronto M5S 3G4, Canada, December 2001. http://www.cs.toronto.edu/ gpenn/ale/files.
5. W. Daelemans, J. Zavrel, A. van den Bosch, and K. van der Sloot. TiMBL: Tilburg Memory-Based Learner Reference Guide, version 4.2. Technical report, P.O. Box 90153, NL-5000 LE, Tilburg, The Netherlands, June 2002.
6. C. P. Rosé and A. Lavie. LCFlex: An Efficient Robust Left-Corner Parser. Technical report, University of Pittsburgh, 1998.
7. E. Brill. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. *Computational Linguistics*, December 1995.
8. http://www.ltg.ed.ac.uk/ mikheev/tagger_demo.html.