Benchmarking Speech Understanding in Service Robotics

Andrea Vanzo[1], Luca Iocchi[1], Daniele Nardi[1], Raphael Memmesheimer[2], Dietrich Paulus[2], Iryna Ivanovska[3], and Gerhard Kraetzschmar[3]

[1] Sapienza University of Rome
{vanzo,iocchi,nardi}@dis.uniroma1.it,
[2] University of Koblenz-Landau
{raphael, paulus}@uni-koblenz.de,
[3] Bonn-Rhein-Sieg University of Applied Sciences
{iryna.ivanovska,gerhard.kraetzschmar}@h-brs.de

Abstract. Speech understanding is a fundamental feature for many applications focused on human-robot interaction. Although many techniques and several services for speech recognition and natural language understanding have been developed in the last years, specific implementation and validation on domestic service robots have not been performed. In this paper, we describe the implementation and the results of a functional benchmark for speech understanding in service robotics that has been developed and tested in the context of different robot competitions: RoboCup@Home, RoCKIn@Home and within the European Robotics League on Service Robots. Different approaches used by the teams in the competitions are presented and the evaluation results obtained in the competitions are discussed.

Keywords: speech recognition, speech understanding, service robots

1 Introduction

Robots are expected to support human activities in everyday scenarios, by interacting with different kinds of users. In particular, domestic robots (i.e. robots operating in our homes) have already entered the market. Examples are cleaning robots, tele-presence robots and assistive robots for elderly care. In these contexts, the interaction with the user plays a key role. For this reason, the importance of enabling untrained users to interact with personal robots has increased. The goal of the research in *Human-Robot Interaction* (HRI) is to realize robotic systems that exhibit a natural and effective interaction with users. Therefore, robots should be provided with sensory systems able to understand and replicate human communication, such as speech, gestures, voice intonation, pragmatic interpretation, and any other non-verbal interaction. Interaction, by definition, requires communication. Humans usually communicate by means of natural language, which can be considered one of the most effective vehicles of interaction. In this respect, the aim of *Human-Robot Interaction in Natural Language* is to develop robots that are able to solve human language references

in the application context they belong. Competition is how humans improve themselves and push forward their abilities. Recently, the robotic research is facing the problem of sharing a common platform and common methodologies to quantitatively compare the different approaches of a particular task.

This paper addresses the problem of benchmarking robot speech understanding through scientific competitions. In particular, we will focus on service robots operating in a home environment and on service robot competitions, i.e., RoboCup@Home[4] and other ones derived by it. The goal of such a benchmark is to measure and evaluate the performances of the speech understanding capability of a general robotic platform, as well as to create a common workspace of discussion on the topic. More specifically, we will describe the *Functional Benchmark on Speech Understanding* (FBM3), performed (in different forms) during the RoCKIn, RoboCup@Home, and European Robotics League Service Robots (ERL-SR) competitions, along with results and configurations of the recent benchmark that took place in the ERL-SR Local Tournament[5] in Peccioli on January 2017. Three teams participated: *(i)* SPQReL team, joint team between Sapienza University of Rome (Italy) and University of Lincoln (UK), *(ii)* b-it-bots@home team at Bonn-Rhein-Sieg University of Applied Sciences (Germany), and *(iii)* homer@uniKoblenz team at University of Koblenz and Landau (Germany).

The paper is structured as follows. In Section 2 we introduce the FBM3, along with the used performance metrics. Section 3 focuses on the different solutions compared in the competition and on the corresponding results. Finally, in Section 4 we analyze these results and draw some conclusions.

2 A Functional benchmark for speech understanding in Service Robotics

This functional benchmark aims at evaluating the ability of a robot to understand speech commands that a user gives in a home environment. A list of commands are selected among the set of predefined recognizable commands, i.e., commands that the robot should be able to recognize within the tasks of the competition or in similar situations. For this competition, the audio files have been randomly extracted from the HuRIC corpus [2], a resource that contains speaker utterances in the home robotic domain. Only commands that meet the requirements of the task have been chosen. Each implemented system is expected to interpret the provided audio files, producing an output according to a suitably defined representation. Such a representation, inspired by the Frame Semantics [8], has to respect a command/arguments structure, where each argument is instantiated according to the arguments of the command evoking verb. It is referred to as *Command Frame Representation* (CFR) (e.g. "go to the living room" will correspond to `MOTION(goal : "living room")`).

[4] http://www.robocupathome.org/
[5] https://sites.google.com/a/dis.uniroma1.it/erl-sr-peccioli/

2.1 Input provided

For the generation of the output, teams are provided with a knowledge base (Frame Knowledge Base, FKB) containing a set of *semantic frames*, in line with [1]. Each frame corresponds to an action that the robot is supposed to perform or, in general, to a robot command. The FKB contains a description of each frame, in terms of allowed arguments (e.g. destination for a *Motion* command), their names and additional information on how to model the activated frame into the CFR. The list of frames and related arguments is the following:

- *Motion*: The action performed by the robot itself of moving from one position to another, occasionally specifying a specific path followed during the motion. The starting point is always taken as the current position of the robot.
  - GOAL: The final position in the space to be occupied at the end of the motion action.
  - PATH: The trajectory followed while performing the motion towards the Goal.
- *Searching*: The action of inspecting an environment or a general location, with the aim of finding a specific entity.
  - THEME: The entity (most of the time an object) to be searched during the searching action.
  - GROUND: The environment or the general location in the space where to search for the Theme.
- *Taking*: The action of removing an entity from one place, so that the entity is in robot possession.
  - THEME: The entity (typically an object) taken through the action.
  - SOURCE: The location occupied by the Theme before the action is performed and from which the Theme is removed.
- *Bringing*: The action of changing the position of an entity in the space from a location to another.
  - THEME: The entity (typically an object), being carried during the bringing action.
  - GOAL: The endpoint of the path along which the carrier (e.g. the robot - and thus the Theme) travels
  - SOURCE: The beginning of the path along which the carrier (e.g. the robot - and thus the Theme) travels

Composition of actions is also possible in the CFR, corresponding to more complex action as the *Pick_and_place* action, represented by a sequence of *Taking* frame followed by a *Bringing* frame (e.g. for the command *"take the box and bring it to the kitchen"*).

2.2 Scoring

During the functional benchmark, different aspects of the speech understanding process will be assessed:

1. The *Word Error Rate* and the *Speech Recognition Accuracy* on the transcription of the user utterances, in order to evaluate the performance of the speech recognition process. While the former counts the errors made in transcribing the speech signal, in terms of words, the latter focuses on the transcription of the whole command.
2. For the generated CFR, the performance of the system is evaluated against the provided *gold standard* version of the CFR, that is conveniently paired with the transcription. Two different performances will be evaluated at this step. One measuring the ability of the system in recognizing the main action, called *Action Classification* (AC), and one related to the recognition of the full command, that is Full Command Recognition (FCR). AC is carried out in term of *Precision*, *Recall* and *F-Measure*, while FCR is measured through Accuracy. For the AC these measures are defined as follow:
   - *Precision*: the percentage of correctly tagged frames among all the frames tagged by the system;
   - *Recall*: the percentage of correctly tagged frames with respect to all the gold standard frames;
   - *F-Measure*: the harmonic mean between Precision and Recall.

   For the FCR, the accuracy is the percentage of correctly interpreted commands.

The final rank of the teams is evaluated considering the FCR. If this score will be the same for two or more teams, the WER will be used as penalty to evaluate the final ranking.

3 Different approaches

The task of robotic spoken commands understanding involves two main subtasks: the speech recognition step, in which a speech audio signal is processed and transcribed, and the natural language understanding phase, in which the actual interpretation of the sentence is extracted. Both of them can be carried out jointly, by following orthogonal approaches or relying on off-the-shelf tools.

3.1 Transcribing Speech

The robustness of Automatic Speech Recognition in domain-specific settings has been addressed in several works. For example, in [12], a joint model of the speech recognition process and language understanding task is proposed. Such model results in a re-ranking framework aiming at modeling aspects of the two tasks at the same time.

There are several works in which the combination of free-form ASR engines and grammar based systems are exploited. In [10] two different ASR systems work together sequentially: the first is grammar-based and it is constrained by the rule definitions, while the second is a free-form ASR, that is not subject to any constraint. Their approach focuses on the acceptance of the results of the first recognizer. In case of rejection, the second recognizer is activated. In [7], a

robust ASR for robotic application is proposed. The system aims at exploiting the combination of a Finite State Grammar (FSG) and an n-gram based ASR to reduce false positive detections. Specifically, a hypothesis produced by the FSG-based decoder is accepted whenever it matches some hypotheses within the n-best list of the n-gram based decoder. A similar approach is the one proposed in [9], where a *multi-pass decoder* is used to overcome the limitations of a single ASR. The FSG is used to produce the most likely hypothesis. Then, the n-gram decoder produces an n-best list of transcriptions. Finally, if the best hypothesis of the FSG decoder matches with at least one transcription among the n-best, then the sentence is accepted. Many off-the-shelf tool for ASR are also available on the market, all of them offering valuable performances in terms of accuracy and usability. *Google Speech API*[6] is probably one the most widespread, due to its availability on mobile devices. The *Microsoft Bing Speech API*[7] offers, along with the speech transcription service, even the voice authentication. *API.AI*[8] allows to recognize the intent of a sentence, that is provided with the speech transcription. The Nuance VoCon[9] is an offline ASR that allows to customize the recognition on the desired domain, by relying on the definition of specific vocabularies and grammars.

3.2 Understanding Robotic Commands

For understanding a robotic command, one of the simplest solution is to rely on keywords or templates that aim at catching the semantic elements for the targeted task [13]. For instance, the *CFR parser* is based on the simple rules defined by the RoCKIn rulebook[10]. First, a verb corresponding to a predefined frame is discovered. Next, the attributes of the frame, e.g., location, object, beneficiary, etc are found. This utilizes lists of predefined values for each attribute. Finally, the results are composed into the required strings and written into a text file. This approach is simple to be implemented and integrated into a robotic architecture. More sophisticated approaches are based on the definition of syntactic grammars, that are often augmented through semantic attachments [4, 5]. However, such approach is limited by the grammar and can be difficult to extend. The solutions that recently receive more consensus within the community of Natural Language Processing are based on statistical Machine Learning and data-driven analysis of the addressed linguistic phenomena [6, 11]. Among them, LU4R[11] [3] - adaptive spoken **L**anguage **U**nderstanding for(**4**) **R**obots - has been developed by the Semantic Analytics Group at the University of Roma Tor Vergata and the LabRoCoCo Group at Sapienza University of Rome. It is a publicly available tool to parse robotic commands, in the context of service Robotics. LU4R is based on the model that has been proposed in [3]. The

[6] https://cloud.google.com/speech/
[7] https://www.microsoft.com/cognitive-services/en-us/speech-api
[8] https://api.ai/
[9] http://www.nuance.co.uk/for-business/speech-recognition-solutions/vocon-hybrid/index.htm
[10] https://www.rockinrobotchallenge.eu/rockin_d2.1.3.pdf
[11] http://sag.art.uniroma2.it/lu4r.html

Spoken Language Understanding (SLU) process is driven by Machine Learning techniques, that allow to generalize the addressed phenomena and improve the robustness against unseen sentences.

Table 1. Results of the FBM on Speech Understanding in Peccioli

Approach	FCR	AC	WER	SR Acc.
GoogleASR+LU4R	69.80%	91.28%	0.06	62.86%
GoogleASR+LU4R+WordHints	60.65%	89.03%	0.05	68.57%
GoogleASR + CFR	40.00%	80.00%	0.09	64.29%
Nuance + LU4R	13.22%	42.98%	0.63	19.64%
MSSpeech + CFR	0.00%	15.52%	0.90	0.00%

4 Conclusion

The comparative performance analysis (reported in Table 1) of different combinations of techniques for speech understanding reported in this paper allows to determine a realistic expected performance in understanding typical commands issued to a domestic service robot.

The results have been obtained by spoken audio acquired during robot competitions, like RoboCup@Home, from different people and in different environmental conditions. These data thus contains all the typical noises affecting robot competitions on service robots and real scenarios. Consequently, the results provide a realistic assessment of typical performance in this task. The presented benchmark for speech understanding in service robotics has been a useful tool for such an evaluation and is available for further comparisons.

The results presented in this paper outline two interesting observations. First, free-form automatic speech recognition systems significantly outperform grammar-based approaches. This is probably due to the difficulty of speech grammars to cover all the possible linguistic phenomena, in terms of lexicon and syntactic rules. In fact, specially when dealing with spoken language, the sentences' structure is often unpredictable. Second, machine learning-based methods for command understanding seem to be more robust and reliable, as they are able to further generalize the lexicon and to cope with possible minor transcription errors. More specifically, the combination between Google ASR and LU4R consistently provided for the best results in several different runs, that are encouraging for its deployment in real situations.

Although results are generally positive, we are still far away from a full understanding of the commands. Future work will include additional studies on this topic in order to further improve the performance and, to this end, we believe that robot competitions and benchmarks and the joint effort of different research groups will significantly contribute to achieve this goal.

Acknowledgement

We want to thank Nuance Communications for sponsoring academic licenses that have been used for the experiments.

References

1. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The berkeley framenet project. In: Proceedings of ACL and COLING. pp. 86–90 (1998)
2. Bastianelli, E., Castellucci, G., Croce, D., Basili, R., Nardi, D.: HuRIC: a human robot interaction corpus. In: Proceedings of the 9th edition of the Language Resources and Evaluation Conference. pp. 4519–4526. Reykjavik, Iceland (may 2014)
3. Bastianelli, E., Croce, D., Vanzo, A., Basili, R., Nardi, D.: A discriminative approach to grounded spoken language understanding in interactive robotics. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016. pp. 2747–2753 (2016), http://www.ijcai.org/Abstract/16/390
4. Bos, J.: Compilation of unification grammars with compositional semantics to speech recognition packages. In: Proceedings of the 19th International Conference on Computational Linguistics - Volume 1. pp. 1–7. COLING '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002), http://dx.doi.org/10.3115/1072228.1072323
5. Bos, J., Oka, T.: A spoken language interface with a mobile robot. Artificial Life and Robotics 11(1), 42–47 (2007)
6. Chen, D.L., Mooney, R.J.: Learning to interpret natural language navigation instructions from observations. In: Proceedings of the 25th AAAI Conference on AI. pp. 859–865 (2011)
7. Doostdar, M., Schiffer, S., Lakemeyer, G.: A Robust Speech Recognition System for Service-Robotics Applications, pp. 1–12. Springer Berlin Heidelberg, Berlin, Heidelberg (2009), http://dx.doi.org/10.1007/978-3-642-02921-9_1
8. Fillmore, C.J.: Frames and the semantics of understanding. Quaderni di Semantica 6(2), 222–254 (1985)
9. Heinrich, S., Wermter, S.: Towards robust speech recognition for human-robot interaction. In: Proceedings of the IROS2011 Workshop on Cognitive Neuroscience Robotics (CNR). pp. 23–28 (September 2011)
10. Levit, M., Chang, S., Buntschuh, B.: Garbage modeling with decoys for a sequential recognition scenario. In: ASRU. pp. 468–473. IEEE (2009), http://dblp.uni-trier.de/db/conf/asru/asru2009.html
11. MacMahon, M., Stankiewicz, B., Kuipers, B.: Walk the talk: connecting language, knowledge, and action in route instructions. In: proceedings of the 21st national conference on Artificial intelligence - Volume 2. pp. 1475–1482. AAAI'06, AAAI Press (2006)
12. Morbini, F., Audhkhasi, K., Artstein, R., Van Segbroeck, M., Sagae, K., Georgiou, P., Traum, D., Narayanan, S.: A reranking approach for recognition and classification of speech input in conversational dialogue systems. In: Spoken Language Technology Workshop (SLT), 2012 IEEE. pp. 49–54 (Dec 2012)
13. Perera, V., Veloso, M.M.: Handling complex commands as service robot task requests. In: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015. pp. 1177–1183 (2015)