IBM Cloud Services enhance automatic cognitive assessment via human-robot interaction

VARRASI, Simone, LUCAS, Alexandr <http://orcid.org/0000-0002-3165-3923>, SORANZO, Alessandro <http://orcid.org/0000-0002-4445-1968>, MCNAMARA, John and DI NUOVO, Alessandro <http://orcid.org/0000-0003-2677-2650>

Citation:

VARRASI, Simone, LUCAS, Alexandr, SORANZO, Alessandro, MCNAMARA, John and DI NUOVO, Alessandro (2018). IBM Cloud Services enhance automatic cognitive assessment via human-robot interaction. In: WENGER, P., CHEVALLEREAU, C., PISLA, D., BLEULER, H. and RODIC, A., (eds.) New trends in medical and service robotics : advances in theory and practice. Mechanisms and Machine Science, 65 . Netherlands, Springer, 165-176. [Book Section]

IBM Cloud Services Enhance Automatic Cognitive Assessment via Human-Robot Interaction

Simone Varrasi [1], Alexander Lucas [1], Alessandro Soranzo [1], John McNamara [2], Alessandro Di Nuovo [1]

[1] Sheffield Hallam University, Sheffield S1 1WB, United Kingdom

[2] IBM Hursley Lab, Winchester SO21 2JN, United Kingdom

Abstract Thanks to recent developments in artificial intelligence and social robotics, Human-Robot Interaction (HRI) can be used as a non-invasive screening tool for the assessment of cognitive decline. In this scenario, the robot manages the assessment by providing the instructions to the patient, registering his/her answers and objectively calculating the final score. This service can help to save time and reach a wider population. From the technical point of view, a challenge is to achieve a highly reliable speech and visual recognition as required for a valid scoring of performance.

In this article, we evaluate a system for cognitive assessment that makes use of the IBM AI Cloud services embodied in one of the most popular platforms for social robotics: the SoftBank Pepper. Results of a pilot study with 16 human participants shows that IBM Cloud services for speech and visual recognition can improve the system performance in comparison with standard interfaces. Importantly, the improvement allows achieving a significant correlation with one of the most used paper-and-pencil tests and, therefore, the study demonstrates the validity of the robotic approach for cognitive assessment.

Introduction

Social robots have been increasingly studied for clinical applications [4]. Human-Robot Interaction (HRI), indeed, is a valid mean to provide patients with valuable services, even in the field of mental healthcare [8]. In the last period, initial evidence has been collected not only on robot-assisted treatments but also on the viability of robotic assessments, for instance in the measurement of Patient Reported Outcome [1] and in early diagnosis of Autistic Spectrum Disorder [2, 7]. When an assessment is managed by a robot, in fact, many advantages allow a more reliable scoring of performances: assessor neutrality, standardization of the interaction, better acceptance of the robotic platform than a non-embodied computer [9], [3]. These features are particularly relevant in the case of a robotic administration of screening tests for early detection of dementia, as robots could guarantee automatic large-scale screening exams for the elderly population.

However, social robots will be able to be integrated into standard evaluation procedures only when the scores provided by them will be valid enough to represent an aid for clinicians, who will be, therefore, supported by the technology for some of their daily tasks. A careful development of robotic interfaces and artificial intelligence is needed to fulfil such a vision. However, in one of our pilot tests on cognitive assessment with the SoftBank humanoid social robot *Pepper* [10], it was found that the score automatically calculated using the robot software often failed to provide the correct score because of the failures of the embedded speech and object recognition interfaces. We concluded that for clinical validity was necessary the revision of the score by a professional supervisor.

This preliminary result prompted this follow-up study on alternative technologies to enhance the reliability of scoring and, therefore, a reliable cognitive assessment that could be used without necessarily involving a professional. An interesting solution is represented by cloud services offered by the major corporations, for instance Microsoft, Google, and IBM. In fact, these cloud services can be considered the state-of-the-art in artificial intelligence and provide a standardized and easily reproducible environment for development and testing of HRI applications. Even if these services were just recently introduced, they were already used for some studies in HRI [6].

After thorough consideration, we selected for further analysis the IBM AI Cloud services ("Watson"), which provide a comprehensive set of easy-to-use tools for speech recognition (speech-to-text) and production (text-to-speech) and object recognition from pictures that met the requirements for our application.

In this paper, we present an exploratory evaluation of the IBM Watson services in comparison with default Pepper's software with the aim to identify which approach can provide an automatic score closer to the one calculated by a professional supervisor, which represents the benchmark in our analysis. In addition, we administered and considered in the comparison the score of a widely used paper-and-pencil test in order to provide also external validation to the robotic instrument.

Materials and methods

Participants

A sample of 16 healthy adults was enrolled in our study (*M*-age = 31.5 years, range = 19-61, *SD* = 14.15). They were all proficient in British English, but only 2 of them were native speakers. The education level was usually higher (*M* = 19.5, *SD* = 4.07).

The robotic assessment procedure (data collection)

The robotic cognitive assessment was inspired by one of the most important and widely used psychometric screening tools for Mild Cognitive Impairment, the Montreal Cognitive Assessment test (MoCA) [5], which is freely available from the official website. We programmed Pepper to administer and score a MoCA-like psychometric assessment, in which the robot provided the instructions. The robotic cognitive test measures the same areas of the MoCAs, hence the subtests have the same names. In practice, speech recognition was used to score: Naming, Attention, Language, Abstraction, Delayed Recall, and Orientation. The visual recognition, instead, was used to assess Visuospatial/Executive skills by asking the user to draw a cube and a clock on two sheets of paper, which were automatically processed by robot's device.

All the administration was audio-recorded by the robot's microphones. The robot took pictures of the user drawings for the visuospatial/executive skills.

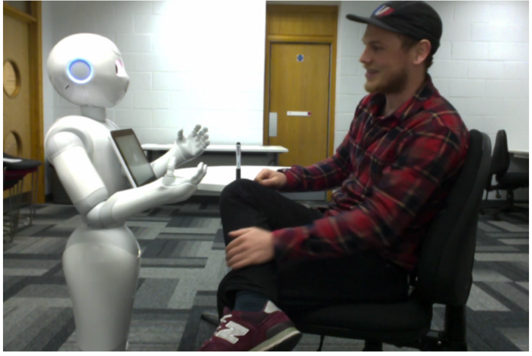

Figure 1. Example of the human-robot interaction during the robotic administration

More detail on the experimental procedure, the assessment protocol, and the robotic software can be found in [10]. The direct result of the robotic administration was twofold: i) a *Pepper score*, calculated by the robot thanks to its embedded AI software, and ii) a *Supervised score*, which was the actual score achieved by the user during the robotic administration, obtained by a professional re-scoring the performance through the recordings available.

Furthermore, in order to provide external validity, the participants were also administered an alternative version of the MoCA test by a human psychologist, who calculated the *Paper and Pencil MoCA score*.

Data processing via IBM Cloud AI Services

The results of the previous pilot experiment [10] show that the automatic processing must be improved, in order to have a reliable score that can be used for automatic screening. To this end, we used the IBM Cloud solutions for post-processing of the data collected from the human-robot interaction, with the aim to establish if these services can be used to improve the automatic assessment.

In particular, we used Watson Speech to Text to perform speech recognition and Watson Visual Recognition to perform the analysis of the hand-drawn images.

Watson Speech to Text

Pepper's default voice recognition system is tailored for recognizing individual words, which is not suitable for some parts of the test where long sentences must be recognized, e.g. the Language assessment. In addition, Pepper's microphones pick up constant noise coming from the cooling fans located inside the head. This has an impact on the quality of the audio source, which in turn negatively affects speech recognition. We planned to overcome these problems by using Watson speech to text service, which makes use of longer sentences for better context analysis and recognition and allows customizing the model to embed noise.

Considering the language background of participants and the characteristics of audios, we used the model *en-GB_BroadbandModel*. Because of the explorative character of this study, we did not set a confidence threshold under which the transcriptions were discarded. For the same reason, we took into account the 10 best alternatives of recognition transcription according to the confidence level.

By processing the audio with standard parameters, we obtained the *basic Watson score*, which is in fact the simplest version.

Also a *customized Watson score* was calculated thanks to the language customization service, which allows training the speech-to-text model for specific recognition requests. Indeed, we created three customized models by adding specific corpora. One model was trained for the recognition of numbers, months and weekdays (Attention and Orientation tasks), one for the recognition of all the words starting with *B*, *F* and *S* (Fluency task), and the last one for all the remaining tasks. The customization weight was set to 0.9.

For both basic and customized, scores were calculated following two scoring approaches: the *exact approach* allowed assigning the points only when the target word had the best confidence level and the string was fully contained in the transcriptions; the *flexible approach*, instead, accepted a certain percentage of error, so that the points were given if the transcription fitted for at least the 70% the target word or string.

Watson Visual Recognition

During the administration, the participants produced hand-drawing of a cube and a clock showing a specific time. They would then present these pictures so that the robot could take a photo. The photos were sent to Watson Visual Recognition for analysis, generating a general class and individual subclass of the object/setting the recognition system deemed as central on the picture. We reprocessed the drawings of cubes and clocks made by participants with the default visual recognition, without any kind of previous training. The system gave a score if the class/subclass contained the object the person was drawing.

The points were given when the cube was recognized as *polyhedron* and the clock as *clock* or *wall clock*. For the hours, points were given if numbers from 1 to 12 were recognized.

Experimental Results

The statistical analyses presented in this section were performed with the SPSS software (version 24). The descriptive statistics were calculated for the various automatic scores (Pepper, Watson), the supervised (benchmark) and the Paper-and-Pencil MoCA score: mean (*M*), standard deviation (*SD*), minimum (*Min*) and maximum (*Max*). The scores are the sum of the subtests scores. Then, Spearman correlations – chosen for the shape of the distribution and the typology of data – were calculated to explore the relationship among each test with the supervised score as well as with the paper and pencil MoCA score, which represents the external validity for the test. Finally, we conducted a repeated measure ANOVA on the differences between each test and the benchmark (supervised score).

Descriptive analyses

The mean Pepper score is 12.69 (Min = 6; Max = 23; *SD* = 4.61), the mean Supervised score is 18.63 (Min = 10; Max =27; *SD* = 4.83), and mean Paper-and-Pencil MoCA score is 25 (Min = 21; Max = 28; *SD* = 2.07).

In the case of Watson, the mean Exact basic score is 10.69 (Min = 2; Max = 17; *SD* = 4.71), the mean Flexible basic score is 11.44 (Min = 3; Max = 20; *SD* = 4.95), the mean Exact customized score is 15.50 (Min = 8; Max = 25; *SD* = 5.16), and the mean Flexible customized score is 16.75 (Min = 9; Max = 25; *SD* = 5.47).

Correlations

Spearman correlations were calculated among the global scores. Results show (Table 1) that the Pepper score does not correlate significantly either with the Supervised score nor with the Paper-and-Pencil MoCA score. The basic Watson score, instead, strongly and significantly correlates with the Supervised score, both in the Exact and Flexible versions. A stronger correlation can be found between the Exact and Flexible customized Watson scores, which also significantly correlates with the Pepper score and the Paper and Pencil score.

TABLE I. SPEARMAN CORRELATIONS (SIGNIFICANT RESULTS IN BOLD)

	Supervised score	Paper and Pencil MoCA score
Pepper score	0.38	0.01
Exact basic Watson	**0.738****	0.476
Flexible basic Watson	**0.737****	0.442
Strict customized Watson	**0.834****	**0.542***
Flexible customized Watson	**0.831****	**0.515***

** *p* < .01; * *p* < .05

Repeated measures ANOVA

A Kolmogorov–Smirnov test for the normality of the raw data showed no significant deviation from normality. A repeated measure ANOVA on the differences between each test and the benchmark revealed a significant effect of the test [F(5,75) = 54.2; p < 0.001]. A post hoc analysis with Bonferroni correction revealed a significant difference between each test and the benchmark (p<0.001) except for the Flexible customized Watson score. Figure 2 shows the departures of each test from the benchmark. As can be seen, the test that deviates the least from the benchmark is the Watson Customized Flex score.

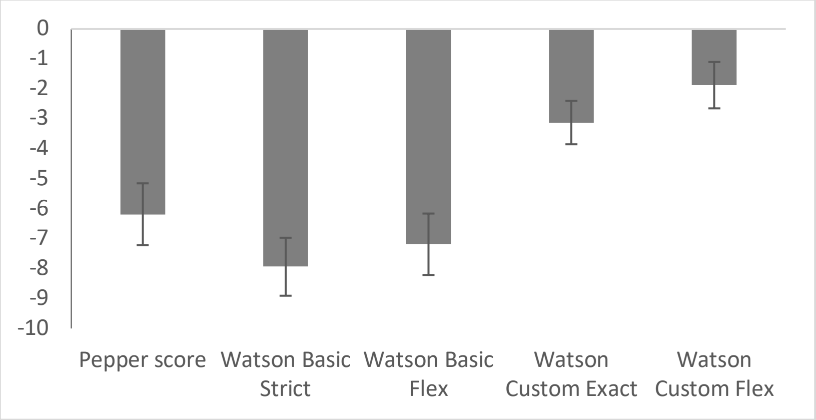

Figure 2. Average departures of each approach from the benchmark.

For a more detailed analysis, Figure 3 presents the average absolute deviations from the benchmark for each subtest. Watson versions strongly improve Visuospatial, Attention and Abstraction, while the basic version struggles with Naming, Delayed Recall and Orientation. However, after the customizations, Watson is always performing better than or as good as the default Pepper algorithms.

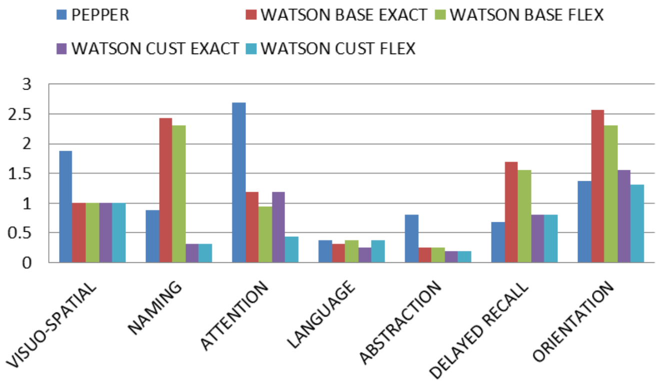

Figure 3. Average Absolute departures of each subtest from the benchmark.

Discussion and Conclusion

In this paper, we presented an exploratory study on the use of IBM Cloud AI services "Watson" for supporting the scoring of a MoCA-inspired psychometric test on a robot. To this end, we collected data from 16 administrations and calculated *Watson scores*: *basic exact*, *basic flexible; customized exact* and *customized flexible*. The Watson scores were compare with the default robot AI, a benchmark *Supervised score* obtained by a professional review of the robot's recordings and the paper-and-pencil MoCA score for testing external validity.

The analysis of the results shows that the score calculated by the robot with its embedded software does not fit our benchmark and lacks external validity. Watson with basic parameters improves the scoring, showing a strong and significant correlation with the *Supervised score*. The best result is obtained with the customized version, which strongly correlates with the *Supervised score* and shows initial clinical validity as it is the only one that significantly correlates with the paper-and-pencil MoCa. The Flexible approach gives the best results by relaxing the confidence thresholds along with the customizations; in fact, this is the only case that doesn't significantly deviate from the benchmark.

We believe that the customizations are made necessary to overcome a structural problem of the Pepper head design, which places the cooling fan close to the rear microphones and the fan noise is always present in the background.

We conclude that the IBM Watson services can contribute in creating a reliable automatic scoring system that can be embedded on a physical device, such as a social robotic platform, and be used for screening of cognitive impairments in order to provide early treatment or for continuous assessment for personalized care. The overall results are very promising and represent a first step towards the development of artificial agents' contribution to psychological assessment.

However, there is still space for improvement and the present study has some of the limitations typical of preliminary studies, including a small sample, which is not balanced because of the majority of not native speakers, and a low quality of audio recordings. Our future work will focus on: i) testing the performance of US Broadband model; ii) further training and audio post-processing to enhance the quality of speech transcriptions; iii) training the visual recognition with different kind of polyhedrons and to reliably assign intermediate points to the clock details.

References

1. Boumans, R. et al.: Proof of Concept of a Social Robot for Patient Reported Outcome Measurements in Elderly Persons. In: HRI'18 Companion: Conference on ACM/IEEE International Conference on Human- Robot Interaction, March 5-8, 2018, Chicago, IL, USA. p. 2 pages (2018).
2. Conti, D. et al.: Robots in education and care of children with developmental disabilities : A study on acceptance by experienced and future professionals. Int. J. Soc. Robot. 9, January, 51–62 (2017).
3. Feingold Polak, R. et al.: Differences between Young and Old Users when Interacting with a Humanoid Robot: A Qualitative Usability Study. In: HRI'18 Companion: Conference on ACM/IEEE International Conference on Human-Robot Interaction, March 5-8, 2018, Chicago, IL, USA. p. 2 pages (2018).
4. Iroju, O. et al.: State Of The Art : A Study of Human-Robot Interaction in Healthcare. I.J. Inf. Eng. Electron. Bus. 9–3, May, 43–55 (2017).
5. Nasreddine, Z.S. et al.: The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. J. Am. Geriatr. Soc. 53, 4, 695–699 (2005).
6. Novoa, J. et al.: DNN-HMM based Automatic Speech Recognition for HRI Scenarios. In: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction. pp. 150–159 ACM (2018).
7. Petric, F., Kovačić, Z.: No data? No problem! Expert System Approach to Designing a POMDP Framework for Robot-assisted ASD Diagnostics. In: HRI '18 Companion: 2018 ACM/IEEE International Conference on Human-Robot Interaction Companion, March 5–8, 2018, Chicago, IL, USA. p. 2 pages (2018).
8. Rabbitt, S.M. et al.: Integrating socially assistive robotics into mental healthcare interventions: Applications and recommendations for expanded use. Clin. Psychol. Rev. 35, 35–46 (2015).
9. Varrasi, S. et al.: A Social Robot for Cognitive Assessment. In: HRI'18 Companion: Conference on ACM/IEEE International Conference on Human-Robot Interaction, March 5-8, 2018, Chicago, IL, USA. p. 2 pages (2018).
10. Varrasi, S. et al.: Social Robots as Psychometric Tools for Cognitive Assessment: a pilot test. In: Springer Proceedings in Advanced Robotics - 10th International Workshop in Human-Friendly Robotics. .