Journal of Intelligent Information Systems (2021) 57:321–345
https://doi.org/10.1007/s10844-021-00648-7

Smart integration of sensors, computer vision and knowledge representation for intelligent monitoring and verbal human-computer interaction

Thanassis Mavropoulos[1] · Spyridon Symeonidis[1] · Athina Tsanousa[1] · Panagiotis Giannakeris[1] · Maria Rousi[1] · Eleni Kamateri[1] · Georgios Meditskos[1] · Konstantinos Ioannidis[1] · Stefanos Vrochidis[1] · Ioannis Kompatsiaris[1]

Received: 10 February 2021 / Revised: 19 May 2021 / Accepted: 20 May 2021 /
Published online: 10 June 2021

Abstract

The details presented in this article revolve around a sophisticated monitoring framework equipped with knowledge representation and computer vision capabilities, that aims to provide innovative solutions and support services in the healthcare sector, with a focus on clinical and non-clinical rehabilitation and care environments for people with mobility problems. In contemporary pervasive systems most modern virtual agents have specific reactions when interacting with humans and usually lack extended dialogue and cognitive competences. The presented tool aims to provide natural human-computer multi-modal interaction via exploitation of state-of-the-art technologies in computer vision, speech recognition and synthesis, knowledge representation, sensor data analysis, and by leveraging prior clinical knowledge and patient history through an intelligent, ontology-driven, dialogue manager with reasoning capabilities, which can also access a web search and retrieval engine module. The framework's main contribution lies in its versatility to combine different technologies, while its inherent capability to monitor patient behaviour allows doctors and caregivers to spend less time collecting patient-related information and focus on healthcare. Moreover, by capitalising on voice, sensor and camera data, it may bolster patients' confidence levels and encourage them to naturally interact with the virtual agent, drastically improving their moral during a recuperation process.

Keywords Human-computer interaction · Sensors · Natural language processing · Pervasive systems · Computer vision · Healthcare

✉ Thanassis Mavropoulos
mavrathan@iti.gr

[1] Centre for Research and Technology-Hellas, Information Technologies Institute, GR 57001 Thermi, Thessaloniki, Greece

1 Introduction

Over the past few years low cost, low energy, wireless devices (cellphones, ambient intelligence (Cook et al. 2009), internet of things (IoT) (Atzori et al. 2010) have surged the consumer market and entered people's routines, assisting in leisure activities with wearables (e.g. biometrics logging), in working environments with ambient sensors (e.g. automated regulation of light/humidity levels) or in various household uses (e.g. dialogue-driven home appliances). The healthcare sector has always been at the forefront of adopting new technologies (Tran et al. 2020; Aouedi et al. 2020; Islam et al. 2015; Tai et al. 2020), since it's a field where technological advancements often have the most meaningful impact, as they directly contribute to the well-being of individuals in dire need of assistance. Patient monitoring based on always-on wireless cameras and remote data analysis offers quasi real-time intervention possibilities that may prove crucial in an emergency situation. Additionally, IoT-infused sensors keep constant track of patient biometrics, while dialogue-competent virtual agents may assist patients by reminding their drug dosage, their exercise program or simply keep them entertained with music/quizzes/news, etc. Moreover, besides patient well-being and satisfaction, health care facilities have also embraced the aforementioned technologies to optimise personnel management, infrastructure operating expenses, and drug/treatment efficacy. Although popular and mature to an adequate degree, the exploitation of these technologies' potential in real-world monitoring, rehabilitation, and care settings has not been manifested to the fullest yet and still presents significant challenges; individual systems have been developed that include interaction with users[1] without however offering a complete and specialised caregiver-to-patient platform.

The present study is an extension of recent work (Mavropoulos et al. 2019) and attempts to tackle those needs by offering a complete health care-oriented solution that comprises an assortment of advanced functionalities, such as multimodal sensor data analysis and management with alarm notifications, a conversational virtual agent with natural user interaction and medical history-respecting dialogue management (DM) that accepts a variety of patient requests, which, to the best of our knowledge, no other system offers in a single platform. Specifically, the proposed platform distances itself from off-the-shelf frameworks by leveraging multimodal verbal (dialogue) and non-verbal (sensor/camera) information management as well as data from patients' medical history to provide a user-friendly experience, using underlying cutting-edge technologies in computer vision, big data analysis, natural language processing and semantics. Notably, the platform: a) is user-agnostic, providing user-requested services to both clinical staff and patients, b) achieves natural user interaction and prolonged engagement, by exploitation of task-oriented and non-task-oriented conversational systems' advantages in an innovative DM framework, capable of social interactions beyond the usual Q&A tasks, c) certifies a holistic approach in patient monitoring, with the deployment of multimodal sensor data (from cameras, sleep sensors, blood sugar/pressure levels and wearable readings) aggregation, analysis and fusion, which to the best of our understanding has not been implemented by other platforms in a single solution, and d) responds to complex queries, by applying reasoning rules to identify the most appropriate topic, since semantics provide a framework which can interconnect data in an intelligent way. A lack of research activity exists in the literature on the topic of combining semantics with DM. Therefore, the system focuses on the benefits of related state-of-the-art research in the form of a smart monitoring framework with dialogue capabilities, which supports

[1] https://medwhat.com/

the physical recuperation and rehabilitation of patients in both clinical and non-clinical environments.

The platform, whose development is part of the research project REA[2] is capable of: a) analysing user input by transcribing his/her voice to text and processing the output for general keywords and named entities of interest in order to understand user needs, b) keeping a history of both user dialogue and status, c) generating an appropriate response using ontology-driven reasoning techniques, based either on information stored in the system's knowledge base (KB) or data retrieved from trustworthy web sources, and d) forwarding the specific response to the user in written or verbal form via text-to-speech techniques. In this work, focus is placed on some of the platform's components that have seen the most advanced development.

Naturally, the advantage such a system provides to involved parties' quality of life, can be two-fold, having both financial and societal consequences; primarily in hospitals, rehabilitation centres and clinics, it makes caregiver-to-patient relations more efficient, by limiting unnecessary interactions. Furthermore, mostly in home environments, it also has direct implications to the patient's emotional state, by minimising the latter's reliance on caregivers and enhancing the sense of independence and self-sustaining. To manage these objectives, the following research directions are being explored: a) the development of a versatile platform that can support both the collection and analysis of verbal and non verbal information, b) the incorporation of multimodal (sensor/camera) data analysis, human-computer dialogue history and patient medical history along with human professional expertise to a KB with reasoning capabilities and c) the deployment of a virtual agent, able to assist medical staff and patients alike.

The rest of this work is organised as follows: An overview of the proposed system's architecture is provided in Section 1.1, while in Section 2 the theoretical background and an outline of the relevant literature are supplied. Section 3 illustrates the experimental framework of our study, and details of the system's potential in a real-life use case scenario are also presented. In Section 4 the experimental results are discussed, while Section 5 presents the initial system evaluation results. Finally, Section 6 concludes the paper and provides future aspirations towards further exploitation of results.

1.1 System overview

Due to the variety of the novel technologies that are being integrated in the framework, a well-defined architecture is required to demonstrate precisely the high-level structural elements of the platform. REA's architecture is illustrated in Fig. 1. It is composed of three different conceptual levels: a) the communication understanding, b) the communication analysis and c) the sensors management. Each level includes a set of software components, hardware infrastructure and data storage implementations. From the considered software components ensemble, we focus in this paper only on the ones that have reached a functional initial version. In the figure, they are shown with green background and bold font in their respective boxes.

All the human-computer interaction is achieved via a mobile device application; all the interactions and the circulated data between the user and the architecture levels of the platform are guided through the *communication understanding* level. This level communicates directly with the user and acts as a central node, permitting data passthrough, and handling

[2] https://rea-project.gr/en/home-en/

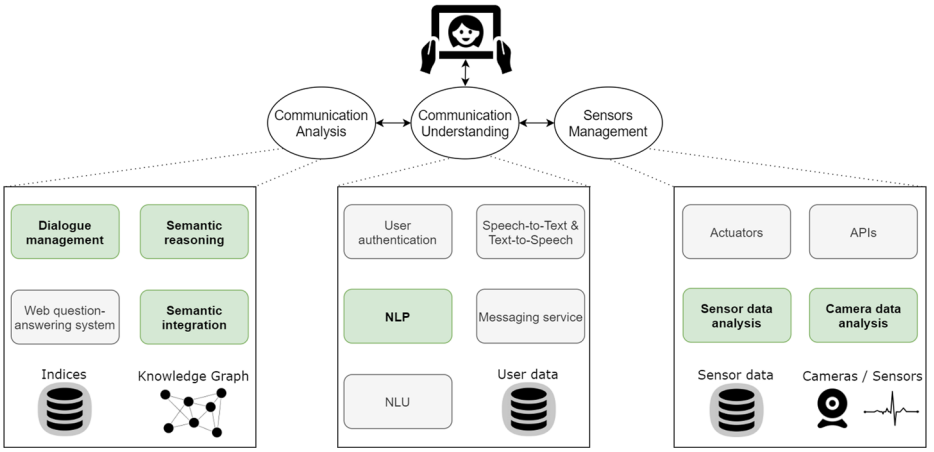

Fig. 1 System architecture. We report progress on the components with green background

all communication between the other levels. It encapsulates components related to primary user interface (UI) actions, such as the authentication for authorised staff, and to the establishment of appropriate communication with the other levels. Also, it performs some initial processing on the user input (speech-to-text, NLP) as well as processing the final system output and converting it into voice format.

The system relies on the *communication understanding* level to carry out essential functionality and on the *communication analysis* one to manage the interconnectivity between levels and their respective components. As such, the system cannot operate optimally without either of them, which is not the case for the *sensor management* level, since, by design, it can serve additional functionality whenever plugged in.

The components that fall into the *sensors management* level are responsible for collecting, and analysing data from various sensors and cameras for patient activity logging and monitoring. The gathered information is either transmitted when a request is dispatched by the *communication understanding* level, or sent proactively through an alert mechanism in the case an emergency is captured by continuous monitoring processes. Apart from just retrieving sensor/camera values, the application programming interfaces (APIs) existing in this level also accept commands that perform actions, e.g. changing the tilt angle of a patient's bed.

The *communication analysis* level is indispensable to the integration of advanced and complicated conversational functionalities into the virtual agent, which improve the system's interaction naturalness. The principal component in this level is the DM module; it maps the latest user utterance into knowledge graphs, leverages the semantics potential (reasoning) to infer content that is not provided explicitly and decides in most cases about the appropriate system response. The decision making process conducted by DM also takes into consideration the data that are generated in the *sensors management* components, retrieved through the *communication understanding* level and stored in the semantic database. Depending on the identified discussion topic, it communicates with semantic-based or Web-based question-answering systems to retrieve information associated with the system response. There are few cases when this level is not advised for the final system response as it can be produced exclusively by the *communication understanding* level. On

these occasions, the *communication understanding* level can send a notification to the DM module to update its dialogue history of the performed system action.

2 Related work

Interactive virtual agents have migrated from consumer products to specialised domains, with health care accommodating many approaches (Savino and Latifi 2019), like avatars that can detect dementia (Tanaka et al. 2017), or others that act as an automated social skills trainer for people with autism spectrum disorders (Tanaka et al. 2017) or ones that promote mental well-being (Ly et al. 2017). While not without issues, Bickmore et al. (2018), Apple, Microsoft, and Amazon in particular have stepped into health care-oriented services (Ravindranath et al. 2018), with Google being also ready to start hospital trials. Health care-centred systems have managed to facilitate patient and caregiver communication (Aiva[3], amazon echo-based) via multipurpose mobile app and are able to support doctors via easy note generators (Suki[4], Robin[5]) or even manage appointment re/scheduling and setting reminders (Merit[6]). The architectures of these platforms have deployed technologies like speech-to-text, text analysis, dialogue management, reasoning mechanisms, language generation, and in REA's case (this work), visual analysis. In the following, we report on the most important works that relate to REA's respective components.

Wearable sensor analysis Fusion is the act of combining data (early fusion) or combining classification results (late fusion) in the field of human activity recognition. A good guide for types of fusion and fusion methods employed in activity recognition studies is provided in Nweke et al. (2019), while in Jain and Kanhangad (2017) concatenation of accelerometer and gyroscope-based vectors was used as early fusion, which outperformed the late fusion method. Deep learning algorithms have also been used as a method to fuse different sensors. In Münzner et al. (2017), convolutional neural networks were utilised to perform early, late and hybrid fusion. The different stages of the algorithm responded to a different type of fusion. Late and hybrid fusion outperformed early fusion. Weights that characterise the performance of a model are also found in late fusion applications. A thorough overview of weighted late fusion techniques can be found in Chernbumroong et al. (2014), where various wearable sensors are combined in order to recognize 13 activities.

Visual analysis Visual analysis is deployed in this paper so as to provide patient activity monitoring capabilities to the system. Related to the task of recognising activities from video streams are popular techniques that utilize depth imagery as well as 3D human skeletons. Most of the earlier works in this domain focused on the extraction of hand-crafted features from depth maps and human joints, in order to characterize their movement (Wang et al. 2012; Xia et al. 2012; Zanfir et al. 2013). More recent studies tend towards leveraging deep learning and specifically Long Short Term Memory (LSTM) networks (Liu et al. 2017). Rhif et al. (2018) extends upon the Lie group representation of Vemulapalli et al. (2014) by involving CNNs and LSTMs. A recent non-deep learning method has been

[3] https://aivahealth.com
[4] https://www.suki.ai
[5] https://www.robinhealthcare.com/
[6] https://merit.ai/

proposed by Luvizon et al. (2017) which encoded multiple spatial and temporal features of different joint subgroups. The aggregation of features is achieved by applying VLAD encoding (Vectors of Locally Aggregated Descriptors) and then optimal feature combinations are found using metric learning. Our work regarding visual analysis is similar to the latter, in its concept of feature aggregation, where we deploy a Fisher encoding schema instead of VLAD.

Natural language processing In our system, a named entity recognition (NER) system is responsible for detecting entities, such as persons, locations and organisations found in user queries and subsequently, either maps the specific query to relevant information stored in the system's KB or activates the information retrieval component. In contrast to past approaches, modern NER systems to operate optimally are neither dependent on feature engineering, nor require specialised resources (Nadeau and Sekine 2007), other than word representations and a small amount of supervised training data (Mikolov et al. 2013; Pennington et al. 2014; Bojanowski et al. 2017), making their usage very versatile and "domain-agnostic". The latest advancements in the field take the form of models based on deep contextualised word representations and an attention architecture called Transformer that learns contextual relations between words. In the latter, the trained vectors consider the input sentence in its entirety and across all layers, instead of just the nearest word context of the top layer found in previous approaches. This approach has improved results even further and newer systems present f1-scores of around 92-93 in the Conll2003 dataset, as can be seen in ELMo (Peters et al. 2018) (92.2 F1), BERT (Devlin et al. 2018) (92.8 F1), Flair (Akbik et al. 2018) (93.09 F1) and in Baevski et al. (2019) (93.5 F1).

Semantics Semantics offer an intelligent interconnection between information coming from heterogeneous sources as they better organise information, limit complexity and extract inferences (Dam et al. 2011) which are very useful in problem solving and decision making systems. Web Ontology Language (OWL) has been designed to represent complex knowledge about entities and relations between them which are saved into RDF triplestores, named Knowledge Bases. Many ontologies have been developed which provide classes and properties to represent different types of models. Web Annotation Data Model (Sanderson et al. 2017) offers the structure to represent an annotation as a set of linked resources, containing a target and a body that is strongly attached to the target. OWL-Time (Hobbs and Pan 2006) ontology describes temporal aspects of resources and the relations between them. Friend Of A Friend (FOAF) (Brickley and Miller 2007) and General User Model Ontology (GUMO) (Heckmann et al. 2005) ontologies cover many different aspects of user profile information. Their usage in REA's population system is shown in Section 3.5.

Dialogue management The main distinction that can be made on conversational systems is the one between task-oriented and non-task oriented systems. The former are developed to perform specific tasks, whereas the latter are utilised in cases where there is no specific goal and when the system's role is to establish a social connection with the user. For example, a task-oriented system can support making reservations in a restaurant (Jurcicek et al. 2011), while a non-task-oriented system can keep company to elderly people (Higashinaka et al. 2014). In this work, a framework that combines task-oriented and non-task-oriented systems has been created. However, such hybrid systems have only recently been studied. In the work of Yu et al. (2017), they apply non-task-oriented strategies when the user's intention is not clear. A work that is closely related to our work is presented in Pragst et al. (2017). In

that agent, a decision mechanism determines whether the conversational agent will rely on the knowledge-based module or on its own data.

3 Methodology

3.1 Datasets

A make or break feature concerning machine learning approaches is the availability of appropriate datasets that will be used to train the relevant computational models. For the wearable sensor analysis application, we used the Heterogeneity Human Activity Recognition (HHAR) dataset (Stisen et al. 2015) that contains accelerometer and gyroscope recordings from smartphones and smartwatches. The HHAR dataset contains devices with different sampling frequencies as its initial use was to model heterogeneity. Six activities were recorded in the dataset: bike, sit, stairs down, stairs up, stand and walk. The visual analysis activity recognition component is trained on the SYSU 3D HOI RGB-D activity dataset (Hu et al. 2017), which contains 480 videos and includes 12 action classes performed by 40 different individuals. The dataset focus is on activities that involve human-object interaction. In the context of the NLP task, the model has been trained and evaluated on the CoNNL2003 dataset (Sang and De Meulder 2003) which includes annotation for four classes of entities: Person (PER), Location (LOC), Organisation (ORG), and Miscellaneous (MISC). The dataset is considered balanced, as it contains similar number of named entity occurrences in the three important categories (PER, ORG, LOC).

3.2 Wearable sensor analysis

In order to leverage the information of many sensors during the project implementation, the research group proposed a late fusion algorithm that utilises weights, a technique that was tested in a previous work (Tsanousa et al. 2019). In the current section the previous applications of the fusion method in Tsanousa et al. (2019, 2020) are extended to one more public dataset of the activity recognition field and the obtained results are compared to the performance of individual sensors' and the results of two other fusion methods.

The human activity recognition process begins with a sliding window with overlap. The sliding windows were taken in order to extract features, similar to Chowdhury et al. (2017). Time domain features (mean, median, maximum, minimum, standard deviation and variance) were extracted from the sliding windows without any further filtering or preprocessing of the data. Several multilabel classification algorithms were tested and the ones that are reported here are: Random Forests (RF), C5 trees and k-Nearest Neighbors (kNN). Each classifier was applied separately to a sensor and the classification results of the algorithms were combined afterwards. For the fusion step, we used the weighted late fusion framework we introduced in Tsanousa et al. (2019), which is based on detection rate, the simple late fusion method of averaging class probabilities and the weighted accuracy method. To derive the weights for the fusion step, the typical steps of a classification framework were applied. An algorithm was trained on the training set and then applied to the test set in order to predict the types of activities. The classification algorithms were applied separately to each sensor and the classification results are afterwards combined using the late fusion methods mentioned. The comparison of the predicted labels with the true labels gives the evaluation metrics that will then be used as weights in the fusion process. Using

the fusion schemes mentioned, we combined two types of sensors, namely accelerometers and gyroscopes. The fusion frameworks are described below:

Weighted accuracy Weighted accuracy (WACC) is a model-based approach, which means that the weights used, enhance the model that has the best performance overall. Accuracy (Metz 2008) is probably the most common evaluation metric used to characterise the performance of a classifier. In this method, the accuracy of a model is divided by the sum of accuracies (Eq. 1). These weights are then multiplied by the respective class probability vectors $P_{ij} = \left\{p_{i1}(x_1), ..., p_{ik}(x_n)\right\}, i = 1, .., m$ and the products of all models were finally added together to create a final class probability vector (Chernbumroong et al. 2014). The class with the maximum probability was assigned to each test case. The formula for weighted accuracy, as described in Chernbumroong et al. (2014), is calculated for each one of the i models.

$$WACC = \frac{Accuracy^{(i)}}{\sum_{i=1}^{m} Accuracy^{(i)}} \tag{1}$$

Detection rate based weighted late fusion This method is considered class-based, thus it pays attention to the recognition of each class, which is usually characterised by the F1-score (Chowdhury et al. 2017) or balanced accuracy. It was introduced by the authors of the current paper in Tsanousa et al. (2019). The supplement of the detection rate of a model ($W_{ij} = 1 - DR_{ij}$) was chosen as weight, believing it will assist the recognition of classes not so easily predicted. The detection rate ($DR = TP/(TP+TN+FP+FN)$) is obtained again in the testing phase of a model, by comparing the true with the predicted labels. The weights are calculated for each model separately and the weighted probability vectors of all sensors are afterwards summed together. The detection rate is different for each class, therefore each class j has different weights. Using ($P_w = W_{ij} P_{ij}$), we multiply the weights with the respective class probabilities and then add the weighted probabilities of the models that will be fused. Again, to assign a class to a test case, we find the class with the maximum fused probability.

Averaging As already stated, each model of those that will be combined, produces k ($j = 1, .., k$) class probability vectors. To combine the results of the different models, we average the respective class probability vectors of the m models combined $\left(P_j = \frac{\sum_{i=1}^{m} P_{ij}}{m}\right)$.

3.3 Visual analysis

Activities in their primitive form can be seen as sequential frames of human pose configurations. We aim to extract low-level pose descriptors, based not only on joint configurations in the spatial domain but also on joint displacement vectors in time the domain. An activity clip can then be characterised by a high-level compact representation, derived by aggregating the information of the full collection of pose descriptors that have been extracted.

Low-level pose descriptors We adopt the original Moving Pose (MP) descriptor by Zanfir et al. (2013) which assumes that the pose, $P(t)$, is a continuous and differentiable function of the body joint positions over time; as such, we can calculate its second-order Taylor approximation in a short temporal window around the current time-step. The first and second order derivatives of the pose function effectively encode information about the temporal changes in pose configuration inside a short temporal window. The final low-level moving

pose descriptor is the concatenation of the static pose vector and the first and second order derivatives. The derivative vectors are then re-scaled to the unit norm, in order to remove irrelevant variation in absolute speed and acceleration across different input sequences. All the train and test skeleton limps are normalised in order to have the same average length between limps of the same type, whilst maintaining the angles between joints.

High-level activity representations Our first contribution lies on the construction of the final activity representations, which are not formed directly by the values of the low-level descriptors themselves, but, rather, from first and second order statistics based upon prototypical descriptors, as explained thoroughly in a previous work of ours (Giannakeris et al. 2020). The process is initiated by reducing the MP descriptor dimensionality using Principal Component Analysis. Next, the aim is to create a statistical model that understands a few prototypical Moving Poses from the training set. To achieve this objective, Gaussian mixtures are used, which are probabilistic models that assume all the data points are generated from mixtures of a finite number of Gaussian distributions (GMM) with unknown parameters. The EM algorithm is applied in order to fit a mixture of Gaussians to the training set and find the optimal parameters. Finally, the full set of low-level descriptors extracted from an activity clip is expressed using the gradients of the log-likelihood of each feature under the GMM, with respect to the GMM parameters. This process is known as Fisher encoding and the final representations are called Fisher Vectors (FV) (Sánchez et al. 2013).

Our previous experiments (Giannakeris et al. 2020) have shown that by aggregating Moving Pose vectors, Fisher encoding can increase recognition performance, compared with the modified-KNN (k Nearest Neighbors) approach of the original Moving Pose paper (Zanfir et al. 2013). Note that the small transitions of pose configurations over short temporal segments have already been encapsulated on the low-level descriptors themselves. However, by aggregating information over the full activity clip, the resulting representations may lack higher-level information about the temporal evolution of prototypical moving poses as an activity is performed. Therefore, our second contribution is to introduce stacked Fisher Vectors extracted from non-overlapping sliding temporal windows over the full activity sequence. To do this a sequence is split into s parts of equal duration and s Fisher Vectors are extracted, as described previously. Finally, all the individual Fisher Vectors are concatenated into a stacked Fisher Vector representation (SFV). During testing, SFVs can now be classified using the inexpensive linear SVM classifier. Both techniques have been evaluated and the results are presented in the following section.

3.4 Natural language processing

In this section, we present the neural network's NER architecture and elaborate on its most important features. Since the main issues that we need to address revolve around the candidate word's form and placement in a given sentence, we employ combined character-level and word-level representation techniques to handle them with efficacy.

The adopted approach is based on a bidirectional LSTM model with a Conditional Random Field (CRF) layer on top. While specific Transformer-based architectures were tested, they were more computationally expensive, which had an immediate impact on the time the model required to run and provide results. Thus, since the system relies on near real-time reporting, such models were not pursued further. LSTMs are ideal when dealing with sequences of data (like text) because of the way the network channels information via its nodes, leveraging its ability to keep information in memory based on history. Any sequential information is being kept in the LSTM's internal state (AKA hidden layer) and is being

updated with each new data via input/output and forget gates. This way the network is capable of predicting the output based on long distance dependencies. The bidirectional nature of the LSTM network manifests with two processes, applicable to each lexical unit of a given sentence that each computes a representation of the lexical unit's left and right context.

Character-level embeddings that take into account the spelling of words are used, to counter issues deriving from language complexity. The technique involves breaking up each lexical unit in its respective characters and then feeding the resulting sequence to a bidirectional LSTM which turns it into a spelling-respecting vector. In order to also respect the syntactic structure of a document and convey each lexical unit's contextual characteristics word-level representations are indispensable. Again, a bidirectional LSTM is used to capture each lexical unit's contextual information (left and right context). Thus, the final word representations combine both of the above embeddings.

Furthermore, to fulfil the task of NER, assigning a NER label to each word in a sentence, the output needs to be annotated accordingly. It has been shown that CRFs (Lafferty et al. 2001) can produce high tagging accuracy; thereby we employ a CRF layer to attribute labels for the whole sentence by leveraging sentence level tag information. The tagging format used follows the BIO scheme (B-TAG for *Beginning of entity*, I-TAG for *Inside of entity*, O-TAG for *Outside of entity*) where each word in a sentence is assigned a label reflecting its role. A named entity frequently spans not just one, but many lexical units and using this format is possible to annotate them efficiently irrespective of their length. As such, in the sentence: *REA, what's the medication that is being administered to Mr. Smith?* the respective annotation reads *B-PER O O O O O O O O O O O O O B-PER O*. This tagging scheme has already yielded promising results, but in consequent testing, the format will be updated to either the BILOU or BIOES variants; they are comparable and usually improve scores even further since they predict dedicated tags for unique/single (U-TAG in BILOU / S-TAG in BIOES) or last/end entities (L-TAG in BILOU / E-TAG in BIOES). Figure 2 highlights the network's structure, which is comprised of an input layer (word embeddings), a hidden layer (Bi-LSTM encoder) and an output layer (CRF layer).

In the specific implementation that is currently being tested, training / validation / test data are comprised of two text files; the first one contains all the user queries that become

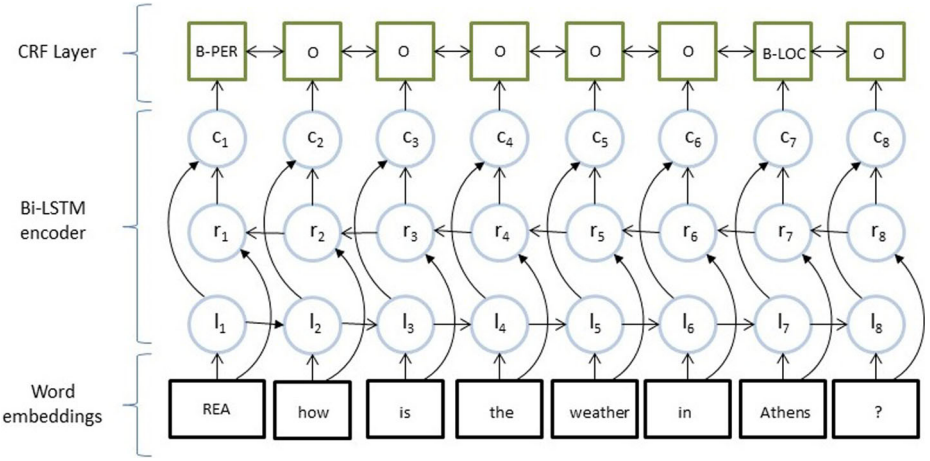

Fig. 2 Bidirectional LSTM-CRF model for Named Entity Recognition

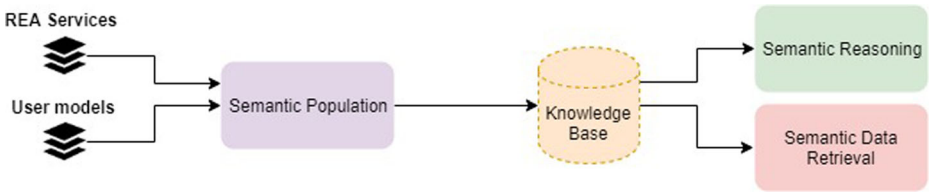

Fig. 3 Semantic Web system architecture

available from the speech-to-text service, while the second contains the BIO annotation of the aforementioned queries, one query per line for both. The results are then outputted in a single file, in a "one sentence token-per-line" format. In the same line, apart from the sentence token the assigned label is also displayed in a tab-delimited format.

3.5 Semantics

We propose a semantic web system which consists of three basic components.

Semantic population system It creates the appropriate representation for metadata coming from heterogeneous sources like verbal communication with the patient, text analysis, camera surveillance, sensor measurements. The system uses OWL 2 ontology language to offer knowledge representation by creating relations between the metadata. GUMO, FOAF and Time ontologies have been extended to meet the needs of the REA project. The semantic population data are saved in the Semantic graph database GraphDB.

Reasoning system It uses Description Logic (DL) services to create important rules for supporting the DM system to infer the discussion topic from the communicated entities. A relevant example reasoning rule is described below:[7]

$$AskTreatmentForInjury \equiv DialogueEntity \sqcap \exists contains.TreatmentReference$$

Each entity that comes from the dialogue with the patient is marked as a dialogueEntity. Text analysis is applied in the concepts stemming from the dialogue procedure. When a treatment reference is found in the concepts, the discussion topic is defined as *AskTreatmentForInjury*.

Semantic data retrieval system It applies semantic queries to the data in the Knowledge Base so to return the most appropriate information according to each user or system request. The semantic data retrieval system is affected by the topic selection of the reasoning system.

The semantic web system's architecture (Fig. 3) is strongly connected with the DM system. The latter provides metadata to the semantic population system, while it communicates with the semantic web system to retrieve the appropriate information to respond to the patient.

3.6 Dialogue management

The Dialogue Management system is based on the hybrid dialogue framework proposed in Kamateri et al. (2019), which consists of five major components: a) the contextual

[7] We use the Description Logic syntax where OWL2 is based on.

modelling and representation, b) the topic detection, c) the dialogue coordination, d) the semantic intelligence and e) the strategy selection. Figure 4 shows the components and the way they interact. In short, the DM system processes the user questions/responses, communicates with the appropriate information retrieval system (semantic-based or Web-based question answering) and decides about the next system response.

Contextual modelling and representation This component represents semantically the verbal and non-verbal input and interconnects it with the existing domain knowledge, regarded as an interface to the semantic web system (Section 3.5). It makes use of two ontologies that comprise the domain and the dialogue model. The domain model stores information that describes an individual's profile, health conditions, received therapy and exercises, and other data related to the individual's behaviour. The dialogue model holds information about the possible conversation topics and the user and system actions.

For the domain model, we made use of two already existing ontologies, extending them based on our application-specific aspects, the COPDology (Ajami and Mcheick 2018) and the IDEF5 (Kultsova et al. 2016), as they best fit our knowledge representation needs. For the dialogue model, we made use of the *move* concept of the OwlSpeak ontology (Ultes and Minker 2014) and we expanded it to map the user and system actions and the supported discussion topics.

The component initially takes as input the key entities, including named entities (Section 3.4), that are associated with the user utterance and extracts resource categories. Then, the resource categories are used to find entities that exist in the domain model by means of the semantic web system (Section 3.5).

Topic detection The topic detection component takes the user utterance and associates it with a discussion topic from the dialogue model. To achieve this, the component can utilise semantic reasoning (Section 3.5) or a simple classification algorithm. In the latter case, the algorithm calculates the probability of the candidate topics, taking into account the communicated entities and domain entities that have already been associated with the candidate topics. The candidate topic set is predefined based on the user requirements and is configured before the deployment of the DM system. Even so, the system is capable of handling any combination of candidate topics without modifying its internal mechanisms (e.g. the topic detection algorithm).

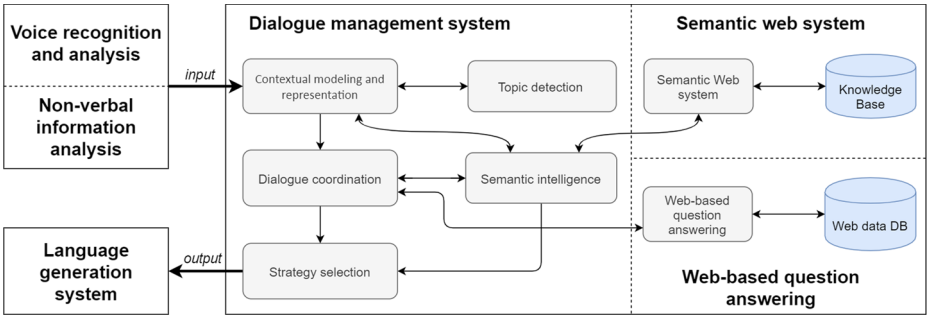

Fig. 4 Dialogue Management system framework

Semantic intelligence The semantic intelligence module updates the domain model with the new information that is circulated during the interaction between the user and the agent. Furthermore, it updates the dialogue history with the extracted entities and the topic of each dialogue turn. Another functionality of this component is to retrieve relevant content from the knowledge base. Last, it applies SPARQL reasoning techniques to generate alerts, reminders and recommendations. These actions are triggered by the knowledge of the preceding discourse in combination with the specific user profile, forwarding system moves that are not directly related to the discussion topic.

Dialogue coordination The dialogue coordination module produces task and non-task oriented system actions and renders the entire component a hybrid one that combines the two different types of conversation systems. Based on the number of identified topics and the matching score, it selects among the following topic-oriented actions: a) predefined topic-based (re-)action, when the matching score of a topic is high enough, b) clarification action, when more than one topics receive an adequate matching score, and c) say-again action, when matching score is quite low.

In order for the system to be able to follow a non-task-oriented functionality, it formulates a set of social-oriented actions: a) switch topic, b) initiate a relevant topic, c) end current topic and open question, d) suggest more information and e) elicit more information. The non-task-oriented functionality is complemented by the actions generated by the semantic intelligence module, including alerts, reminders and recommendations.

Strategy selection The role of the strategy selection component is to determine which candidate action is the most appropriate to be returned as a system response. Three different mechanisms have been considered for the REA project: a) Random selection, this is a simplistic approach selecting randomly between the chosen topic-oriented action and the extracted actions from the semantic intelligence module. In the cases when the topic-oriented approach fails to return an adequate score, all the social-oriented actions are also considered in this random pick of system action. b) *Selection based on the dialogue context*, where the extracted candidate responses are ranked using significance scores and the top-ranked one is returned to the user (e.g., responses generated by the semantic intelligence module are preferred over topic-oriented actions), c) *Selection based on reinforcement learning*, where a simplified version of the Q-learning approach, presented in Yu et al. (2016), can be utilised to train a policy for decision making. The reward function parameters are defined with the help of domain experts' knowledge and comprise the turn index, the number of times each strategy was selected and the most recently used strategy.

The current version of the system supports the first two of the three strategy selection methodologies. The reinforcement learning approach is still under design and more parameters may be added in the reward function when it is finalised.

3.7 Platform deployment and use case example

This section illuminates REA's deployment, installation and user evaluation phases and settings. Then, a use case example is presented to showcase the utility that a health-oriented virtual agent may offer.

Table 1 Doctor-agent use case scenario

Actor	Interaction
(i1) Doctor	REA, did Mr. Goines leave his bed today?
(i2) REA	Yes, he walked for 10 minutes, was sitting for 30 minutes and visited the restroom 3 times.
(i3) Doctor	Did he have fever during visiting hours?
(i4) REA	No, his body temperature was 36.6 degrees Celsius.
(i5) Doctor	What is the weather tomorrow?
(i6) REA	It will be sunny, around 21 degrees Celsius.
(i7) Doctor	Please remind the patient to have a walk at noon.
(i8) REA	Reminder set for tomorrow at noon.
(i9) Doctor	Thank you, that would be all.
(i10) REA	Have a nice day Dr. House!

3.7.1 Evaluation environments

System deployment and installation will be carried out in three phases. During the first two phases, REA will be installed (first phase) and then evaluated (second phase) in a clinical setting. This process is expected to yield useful feedback from patients and medical professionals in order to report the platform's shortcomings and update the system accordingly in view of the next development iteration. The user partner that will host the REA system and partake in its assessment is the EVEXIA[8] rehabilitation centre, which is an established health care service provider in the area of Thessaloniki for the past twenty years. The system will be initially installed to monitor two single-bed hospital rooms, with a plan to add another eight during the second and third phase, for a total of ten rooms. The installed equipment will include all the required sensors and cameras, which will monitor patient behaviour. Finally, a home environment installation (two residences) will take place during the third phase that will feature the updated version of the platform.

3.7.2 A use case scenario

In an effort to convey the level of assistance REA is capable of providing in a hospital environment, a simple use-case interaction between a clinician and the system is presented. The platform's multifarious functionality is highlighted by placing emphasis on potential reminders—alerts that REA can facilitate—and on the informational needs that it can accommodate. In Table 1, the example dialogue between a doctor and REA is presented, while the involvement of all the system components in the human-machine interaction is broken down in the subsections that follow.

Language understanding It follows a three-step procedure: first, it transforms the verbal communication information into text, then it detects concepts, named entities and the relations between them, and in the end, it identifies the BabelNet synset of each term. For instance, in sentence (i1) the concepts "Goines", "leave", "bed" and "today" are extracted, while in (i3) the detected concepts are "fever", "visiting" and "hours". Each concept is matched with a BabelNet synset URL e.g. fever corresponds to "https://babelnet.org/

[8] http://www.evexia.com/en/

synset?word=bn:00033883n&lang=EN&langTrans=EN". The patient's profile is accessed by exploiting the extracted named entity "Goines".

Reasoning After the concepts' recognition by the language understanding step, a reasoning mechanism is applied to detect the semantic meaning behind the concepts. In sentence (i3), for example, the mechanism detects a fever context when the measurement of the patient's body temperature is in a range from 37°C to 42°C, as shown in the following example. In case that the constraint is not satisfied, the patient has a normal body temperature.

$$FeverContext \equiv PatientHealthMeasurements \sqcap \exists hasCondition.(BodyTemperatureRange \sqcap \exists value.high)$$

where *high* is the symbolic representation of high temperature, assigned to the person when the measurements indicate fever ($temperature > 37$).

Dialogue management It is involved in all interaction turns in this dialogue. Firstly, it discovers the discussion topic and the main entities of the user utterance. For example, in (i1) the topic is *AskPatientMovementForSpecificDay* and in (i5) it is *AskWeatherForSpecificTime*. Indicative entity categories identified are temporal (*today* in (i1) and *visiting hours* in (i3)), or named entities (*Mr. Goines* in (i1)). In addition, using semantic knowledge it manages to map pronouns like *he* in (i3) to the correct patient.

Based on the identified topic, it generates candidate actions that may retrieve information from the Knowledge Bases, the sensors/cameras data and the Web-based question answering system. For answering (i1), it consults semantic knowledge to find actions related to leaving from bed and sensors/camera data for the details of these actions, while for (i5) it redirects the user request into the Web-based question answering service.

Apart from predefined topic-based reactions, it can generate actions which are based on conversational history. A typical one is an *alert* that can be triggered if a user repeats the same question many times. Finally, it is responsible for producing social oriented actions like reminders (i7).

Visual analysis The monitoring functions of the visual analysis components provide information about the activities Mr. Goines was doing this particular day, as well as the rooms he was detected by the cameras to walk into, i.e. in this case the bathroom (i2).

Wearable sensor analysis Wearable sensors are involved in interaction (i2) to recognise the activities performed by the patient. The recordings of the patient's wearable sensors are extracted, along with the respective timestamps, which are used to estimate the duration of the activities. These recordings are analysed and using a classification algorithm, the performed activity is recognized. The timestamps are used to estimate the duration of the activities.

4 Components evaluation and results

In this section, results of the aforementioned components are reported. The presentation is limited to the subsystems where a technical evaluation is possible, based on current development.

4.1 Wearable sensor analysis

Having already evaluated our method in certain public datasets, and in order to check its versatility, we test its performance on one more public dataset, the HHAR, to combine previous efforts with accelerometers and gyroscopes. The results of the proposed method are compared with the individual performance of sensors and with two other late fusion methods, the weighted accuracy and averaging. We subset the initial data based on the model of the device so that all recordings have the same frequency. Here we report the results for the Samsung s3 mini model, with a frequency equal to 100 Hz and for the LG Nexus 4 with 200 Hz sampling frequency. The time domain features are extracted from a sliding window of 2 seconds with 1 second overlap, which in the S3 mini subset responds to 200 recordings with 100 overlap, since the sampling frequency is 100 Hz, while in the LG Nexus 4 subset respond to 400 recordings with 200 overlapping. Thus the time domain features mentioned earlier are calculated from the recordings that constitute the time window.

Table 2 presents the accuracy values of the classifiers applied to the recordings obtained from each device. In both devices, for all classifiers, fusion of the results of two sensors outperforms individual performances. The values of the three fusion methods are quite close and do not show remarkable deviations.

This application confirmed the importance of late fusion in the improvement of recognition rate of activities when using wearable sensors. It is also obvious that our method is quite promising and performs equally well with other fusion methods.

4.2 Visual analysis

Until now, we have tested the Moving Pose descriptor with Fisher encoding using a single FV to two well-known RGB-D (RGB and Depth) datasets for activity recognition (Giannakeris et al. 2020). In this work, we extend our experimental efforts to another RGB-D activity dataset so as to examine the applicability of SVF for various values of s and several GMM vocabulary sizes. The SYSU 3D HOI dataset (Hu et al. 2017) poses several key challenges which are different from the other two datasets previously examined: a) There are similarities between the manipulated objects among some activities (e.g. sweeping and mopping); b) there are many subjects performing the activities, therefore more inter-subject variations can be observed for the same type of activities due to the different characteristics of participants. We follow the cross-subject splitting approach for evaluation, where half the subjects are used for training and the other half are used for testing. We present the mean accuracy of 30 random cross-subject splits, which is the designated evaluation protocol for this dataset (Hu et al. 2017).

Table 2 Accuracy values for the Samsung S3 mini subset and the LG Nexus 4 subset

		Accelerometer	Gyroscope	DR fusion	Averaging	WACC
	RF	0.7225	0.6352	0.8696	0.8709	0.8593
S3	C5	0.7013	0.5934	0.8131	0.8144	0.8189
	kNN	0.7887	0.5768	0.8234	0.8491	0.8343
	RF	0.6776	0.6426	0.7977	0.7948	0.7948
Nexus	C5	0.6928	0.6148	0.7727	0.774	0.7747
	kNN	0.6963	0.6131	0.7203	0.7621	0.7582

We experiment with 3 different values for the number of stacked Fisher Vectors, s, where $s = 1$ is essentially a single Fisher Vector encoding the full sequence. We also experiment with 4 different vocabulary sizes for the GMM. Table 3 shows the results after 30-fold cross-validation of all the combinations of s and Number of Gaussians. The color map indicates lower accuracy with yellow hues and higher accuracy with orange hues. The accuracy of a single Fisher Vector (bottom row) is low regardless of the vocabulary size. This means that even if the number of prototypical poses in the model is increased, the method still cannot outperform the variant where multiple Fisher Vectors are stacked. Instead, the best results are obtained when 32 prototypical poses are trained and 3 Fisher Vectors are stacked. Performance slightly drops in the GMM axis after this point. This may be due to redundancies in the increased set of prototypical poses that are created for higher values. Note that these parameters are chosen to optimise performance for this particular dataset. In a larger scale dataset where a lot of activity categories may exist, the method is expected to benefit from a bigger GMM vocabulary that can capture higher variance. In addition, when dealing with longer activity sequences, stacking more Fisher Vectors should also yield better performance. Therefore, those parameters introduce flexibility to deal with various training data characteristics, with regards to both the number of samples as well as the duration of the clips. This is especially important, since the final system will be trained in a continuously increasing pool of data as more and more patients enter the program. Regarding efficiency levels, our proposed method achieves fast inference times with processing speeds of 300-400 frames per second, which is enough for real-time monitoring of ten patients. All the experiments were performed using a quad-core CPU clocked at 3.50GHz with 32GB of available RAM.

Table 4 shows the comparison with other related works on this dataset that use skeleton based features (all use 30-fold cross-validation). Our Moving Pose + SFV methodology achieves higher recognition accuracy compared to LAFF skeleton features of Hu et al. (2016) and the ST-LSTM of Liu et al. (2017). The first uses a 3-level pyramid to capture temporal structure, while the second learns long-term dependencies in sequential skeleton data using spatio-temporal LSTMs. Both works can process videos real-time, but the second reports inference times using a GPU. On the contrary, Dynamic Skeletons (Hu et al. 2017) surpass our method in recognition accuracy by 1.1%, using temporal pyramid Fourier features. Our method is also surpassed by a deep progressive reinforcement learning framework

Table 3 Mean accuracy (%) for different GMM + SFV parameters

Number of stacked FVs \ Number of Gaussians	16	32	64	128
5	73.07	74.08	72.84	73.24
3	72.66	74.35	73.08	72.62
1	66.73	67.44	67.65	67.27

Table 4 Comparison with State-of-the-art

Method	Accuracy (%)
LAFF (SKL) (Hu et al. 2016)	54.2
ST-LSTM (Tree) (Liu et al. 2017)	73.4
Dynamic Skeletons (Hu et al. 2017)	75.5
DPRL + GCNN (Tang et al. 2018)	76.9
Moving Pose + SFV	74.4

for key frame selection, which uses a graph-based convolutional neural network to model dependency between human joints (Tang et al. 2018). DPRL + GCNN has a higher performance by 2.5%, but it also heavily relies on deep learning which comes along with the requirement of GPU deployment in order to achieve fast inference times. Overall, our method's robustness level is close to the current state-of-the-art results—at least in the examined dataset—whilst supporting real-time processing of approximately ten video streams. Note that, the activity recognition function of REA, is also backed up by wearable sensor analysis, as already discussed, therefore possible shortcomings of either component may be covered by the other. However, such complementary characteristics have not yet been thoroughly investigated yet. In case of conflicting results from multiple sensors/cameras we follow the simple route of accepting the most confident classifier. In general, for a multi-modal scenario, it is best to use fusion techniques that combine the classification probabilities, instead of the predicted classification labels. Furthermore, there are more elegant approaches for fusion like stacking and bagging that train the classifier on the classification probabilities of the two sources, however this has not been tested yet on a real-time scenario.

4.3 Natural language processing

Network parameters and training The parameters that were used during training of the Bi-LSTM-CRF models are displayed on Table 5. The word representations used and evaluated as input in the current model are the publicly available, pre-trained 300-dimensional GloVE embeddings (Pennington et al. 2014) trained on the common crawl corpus for the English language, the ELMo embeddings (Peters et al. 2018), and the BERT embeddings (Devlin et al. 2018). In the current testing phase, the same settings will be applied to both English and Greek (when available). However, in future phases, these will be fine-tuned to perform optimally, taking advantage of all differentiating features of the specific resources.

Table 5 Bi-LSTM-CRF settings used for the NER task

Training parameters	Value
Optimiser	Adam
Character embeddings dimensions	100
Word embeddings dimensions	300
Dropout rate	0.5
Epochs	25
Batch size	20
LSTM size	100

Table 6 NER performance of the proposed system vs. other popular approaches

System (CoNLL2003)	Precision	Recall	F1-score
Our system (ELMo embeddings)	91.63	93.01	92.32
Our system (BERT embeddings)	91.46	92.32	91.88
Lample et al. 2016	90.95	90.94	90.97
Best shared task system: Florian et al. 2003	88.99	88.54	88.76
Baevski et al. 2019	(not reported)	(not reported)	93.5

Results To evaluate system performance the values for precision, recall and F1-score measures were computed. At this point the results concerning the English language are very similar to the state-of-the-art, yielding an F1-score of ˜92 (Table 6). To help improve this score, on future iterations the model will be updated with newer embeddings (e.g. Flair Akbik et al. 2018) and annotation (to the BILOU/BIOES format).

Next, we introduce four example tweets extracted directly from the REA platform after having been processed by the NER tool. Integrated NER annotation is included to better convey the application process, while in Table 7 the difference in the results when using different embeddings is illustrated.

1. What events are available in **Athens** (LOC) today?
2. I wish to know more about the **Brexit** (MISC).
3. I wish to hear the latest **Coldplay** (MISC) album.
4. **Rea** (PER), what medication is being administered to Mr. **Smith** (PER)?

5 End-user evaluation

Towards assessing the performance of the entire REA platform, we are conducting a user-oriented evaluation process. The principal reason is that regardless of the efficiency of the individual components, based on objective benchmarking metrics, the actual evaluation decision is in the hands of the end-users.

The aim of the initial user-centred assessment is to assess users' thoughts and experiences when using the system to support its development. Despite some functionalities of REA are not yet finalised, the first prototype of the system with limited features was released and available to test and evaluate in a real environments. The initial version of the system included only a subset of the outlined deployed options, however, feedback was crucial and defined what the platform progress should be towards the final phase.

Table 7 Performance evaluation of the proposed system with different embeddings

System configuration	Correct entities	Wrong entities
Current embeddings (Glove)	Athens, Smith	Brexit (ORG), Coldplay (LOC), Rea (ORG)
With ELMo	Athens, Brexit, Smith	Coldplay (ORG), Rea (ORG)
With BERT	Athens, Brexit, Rea, Smith	Coldplay (ORG)

The selected environment for the evaluation activity was the EVEXIA rehabilitation centre. The first pilot was set up in two beds of the clinic, where respective patients interacted with the system and provided feedback. More specifically, the user groups in this assessment procedure were as follows:

– 20 patients that were in their rehabilitation phase.
– 15 individuals from the clinic personnel with diverse professional backgrounds (4 doctors, 3 nurses, 3 physiotherapists and 5 caretakers of other specialties).

It should be noted that the software has been designed to execute real-time requests from multiple different users/patients.The number of different users only depends on the available hardware resources. Any modern multithreaded server with a reasonable amount of system RAM is adequate to handle the REA system, since none of the developed sub-systems have high memory complexity or require excessive amounts of system memory by themselves. The central component of the system is occupying approximately 50MB of RAM memory, while the rest of the components occupy 20GB of system RAM. During the evaluation phase, REA's capacity for simultaneous monitoring was up to 2 patients. Accordingly, the evaluation for 20 patients was carried out during 10 different cycles.

A three-stage evaluation was organised, with each stage lasting three days. Three different methods were employed for obtaining the required results: a) observation by the medical staff during the pilots, b) interviews with questions aiming to elicit qualitative information about the platform, c) structured questionnaires. For the latter method two types of questionnaires were prepared, for the patients and the clinical personnel respectively, so as to address different aspects of the system. Specifically, the questionnares' structure and overall design was driven by the user requirements, which were stated at the beginning of the project, and included user interface satisfaction (QUIS-short version questionnaire, Chin et al. 1988) and ease of use (PUEU questionnaire, Davis 1989) focused questions. Thus, various quality parameters were used to compile the questions that accompanied the evaluation process and attempted to address these needs. The questionnaire questions revolved around some qualitative variables such as the degree of user satisfaction with the overall functionality of the system, the separate services provided, the system's response speed, the variety found in the different types of interaction with the system, the satisfaction of user needs and requirements, and the usability and user friendliness, among others.

To address these needs the questions asked to the patients were mainly designed to measure the value of the implemented functionalities and the usability of the virtual agent (e.g. "Is REA returning information pertinent to the questions?" - Table 8), while the questions directed to the clinical workers were designed to capture the capability of the software to increase the efficiency of their efforts in providing care-giving services. They answered a Likert scale questionnaire (1-strongly agree, 5-strongly disagree), which the consortium prepared and reviewed in two separate sessions. Each question followed the original Likert weighting scale, while the average score for the questions was 4.26.

The provided input from all users of REA was gathered to formulate a cumulative conclusion with respect to their experience during the tests. The general feedback from both patients and the clinic staff was positive, along with some negative points regarding specific aspects of the platform that will be taken into consideration in the next development phase. Patients were considerably satisfied with the set of supported functionalities, the tool's responsiveness and the user friendliness of the system. The majority of them stated that the services provided were adequate and there were no missing features. The system was fulfilling their needs in fairly quick response times. The main negative element that was

Table 8 Sample of the questionnaire used for patient evaluation

Please rate your agreement on the following statements:						
REA provides the right amount of information.						
strongly agree	O	O	O	O	O	strongly disagree
REA provides very little information.						
strongly agree	O	O	O	O	O	strongly disagree
REA provides excessive information.						
strongly agree	O	O	O	O	O	strongly disagree
REA clearly communicates its intention.						
strongly agree	O	O	O	O	O	strongly disagree
I always understand what REA is telling me.						
strongly agree	O	O	O	O	O	strongly disagree
REA understood what I asked for.						
strongly agree	O	O	O	O	O	strongly disagree
I got the information I wanted.						
strongly agree	O	O	O	O	O	strongly disagree
REA returns contradictory information.						
strongly agree	O	O	O	O	O	strongly disagree
REA returns information pertinent to the questions.						
strongly agree	O	O	O	O	O	strongly disagree
REA works the way I expected.						
strongly agree	O	O	O	O	O	strongly disagree
REA needs too much time to respond.						
strongly agree	O	O	O	O	O	strongly disagree

noted was the inability to properly entertain the users and reinforce their engagement in the system. The addition of more social actions in the virtual agent will serve in mitigating this issue. Apart from that, the need to extend the platform's knowledge about various topics was stressed out, as it was unable to answer some of the users' questions.

The professional clinic workers were equally affirmative about REA's potential. In particular, some of the advantages that were observed from this group of users include the system's effectiveness, consistency, simplicity and reliability. They were also impressed by the swiftness of the agent's replies. Their proposals for improving the service quality of REA focused on two basic concerns. The first recommendation stated that the amount of available information provided by the system should be enriched to assist their awareness about additional topics (e.g. the medication of a patient). The second issue was, similarly to the patients' feedback, the level of entertainment REA was competent to provide to the patients. This aspect was highlighted as important by the care-giving professionals, since the rehabilitation process in a clinic environment can be stressful at times and positive emotions have respective impact on patient state (Richman et al. 2005). Overall, regardless of the reported downsides, the users that participated were excited about the prospect of integrating a platform like REA inside a clinic environment.

6 Conclusion and discussion

The work that has been presented includes development details for the main modules of a patient-monitoring framework designed to assist medical stuff and patients in clinical or home environments during the rehabilitation process. To achieve acceptance, the system should provide solutions to users' everyday routine, assisting in otherwise time-consuming tasks; via REA's utilisation, caregivers would be able to minimise mundane tasks, which will be largely automated, while patients' morale and self-esteem would be increased, by handling themselves previously unattainable tasks. Even trivial ones, like regulating room temperature or switching on the lights, which would have been impossible without help. Advances in sensor data analysis, computer vision, dialogue management, natural language processing and semantics have been combined to achieve natural and seamless human-computer interaction.

Specifically for computer vision, an activity recognition framework from depth cameras has been proposed in order to understand the activities that take place. The identification of meaningful activities has been completed before the training of the final activity classification models, according to the user requirements provided by the clinical user-partner. However, if the need to include a more diverse set of activities arises, then the classifiers for each module should be retrained with the updated set. At a later stage, we will also deal with activity detection. Moreover, fusing the visual and sensor modalities is currently on our future plans, so as to provide a unified output to the system. Likewise, the framework of the DM, the component that is responsible for handling most cases of the user-agent conversation has also been established. Next, the DM subsystem is going to be evaluated to measure its performance and proceed with adaptations wherever necessary (including the completion of the reinforcement learning module). As regards to language understanding, a NER subsystem is responsible for retrieving proper name-related user requests, while a semantic KB collects all data and applies semantic reasoning to enrich the results and support the DM system. Applying more complex reasoning rules and updating the ontologies to represent additional information are left as future work for the semantic system, while extending the NER system to handle Greek constitutes a high-priority future prospect.

To ascertain that the current work is satisfactory and within scope of the initial objectives, meticulous trials have been scheduled in regular prototype-testing periods, based on the implementation of key indicators that will measure the platform's impact. A first evaluation has taken place, but further tests have been suspended because of the obstacles introduced by the global pandemic (COVID-19 outburst). As might be expected, the user-centered evaluating process is based on user requirements which emanated from both caregivers and patients in the project's preparation phase. The same users provided initial feedback on system utility, user adoption and satisfaction in order to attain early detection and mitigation of deterrent factors and conditions.

Acknowledgements This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship & Innovation, under the call RESEARCH-CREATE-INNOVATE (project code: T1EDK-00686).

Declarations

Conflict of Interests The authors declare that they have no conflict of interest.

References

Ajami, H., & Mcheick, H. (2018). Ontology-based model to support ubiquitous healthcare systems for copd patients. *Electronics*, *7*(12), 371.

Akbik, A., Blythe, D., Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1638–1649).

Aouedi, O., Tobji, M.A.B., Abraham, A. (2020). Internet of things and ambient intelligence for mobile health monitoring: A review of a decade of research.

Atzori, L., Iera, A., Morabito, G. (2010). The internet of things: A survey. *Computer Networks*, *54*(15), 2787–2805.

Baevski, A., Edunov, S., Liu, Y., Zettlemoyer, L., Auli, M. (2019). Cloze-driven pretraining of self-attention networks. arXiv:1903.07785.

Bickmore, T.W., Trinh, H., Olafsson, S., O'Leary, T.K., Asadi, R., Rickles, N.M., Cruz, R. (2018). Patient and consumer safety risks when using conversational assistants for medical information: an observational study of siri, alexa, and google assistant. *Journal of medical Internet research*, *20*(9), e11510.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Brickley, D., & Miller, L. (2007). *Foaf vocabulary specification 0.91*. Citeseer.

Chernbumroong, S., Cang, S., Yu, H. (2014). Genetic algorithm-based classifiers fusion for multisensor activity recognition of elderly people. *IEEE Journal of Biomedical and Health Informatics*, *19*(1), 282–289.

Chin, J.P., Diehl, V.A., Norman, K.L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 213–218).

Chowdhury, A.K., Tjondronegoro, D., Chandran, V., Trost, S.G. (2017). Physical activity recognition using posterior-adapted class-based fusion of multi-accelerometers data. *IEEE Journal of Biomedical and Health Informatics* (99), 1–1.

Cook, D.J., Augusto, J.C., Jakkula, V.R. (2009). Ambient intelligence: Technologies, applications, and opportunities. *Pervasive and Mobile Computing*, *5*(4), 277–298.

Dam, H.V., Engberg, J., Gerzymisch-Arbogast, H. (2011). *Knowledge systems and translation* Vol. 7. Berlin: Walter de Gruyter.

Davis, F.D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 319–340.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.

Giannakeris, P., Meditskos, G., Avgerinakis, K., Vrochidis, S., Kompatsiaris, I. (2020). Real-time recognition of daily actions based on 3d joint movements and fisher encoding. In *International Conference on Multimedia modeling, 5-8 January 2020*: Springer.

Heckmann, D., Schwartz, T., Brandherm, B., Schmitz, M., von Wilamowitz-Moellendorff, M. (2005). Gumo–the general user model ontology. In *International Conference on User Modeling* (pp. 428–432): Springer.

Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., Matsuo, Y. (2014). Towards an open-domain conversational system fully based on natural language processing. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 928–939).

Hobbs, J.R., & Pan, F. (2006). Time ontology in owl. *W3C Working Draft*, *27*, 133.

Hu, J.F., Zheng, W.S., Lai, J., Zhang, J. (2017). Jointly learning heterogeneous features for rgb-d activity recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(11), 2186–2200.

Hu, J.-F., Zheng, W.-S., Ma, L., Wang, G., Lai, J. (2016). Real-time rgb-d activity prediction by soft regression. In *European Conference on Computer Vision* (pp. 280–296): Springer.

Islam, S.M.R., Kwak, D., Kabir, M.D.H., Hossain, M., Kwak, K.-S. (2015). The internet of things for health care: a comprehensive survey. *IEEE Access*, *3*, 678–708.

Jain, A., & Kanhangad, V. (2017). Human activity classification in smartphones using accelerometer and gyroscope sensors. *IEEE Sensors Journal*, *18*(3), 1169–1177.

Jurcicek, F., Keizer, S., Gašić, M., Mairesse, F., Thomson, B., Yu, K., Young, S. (2011). Real user evaluation of spoken dialogue systems using amazon mechanical turk. In *Twelfth Annual Conference of the International Speech Communication Association*.

Kamateri, E., Meditskos, G., Symeonidis, S., Vrochidis, S., Kompatsiaris, I., Minker, W. (2019). Knowledge-based intelligence and strategy learning for personalised virtual assistance in the healthcare domain. In *Proceedings of Semantic Technologies for Healthcare and Accessibility Applications (SyMpATHY)*.

Kultsova, M., Potseluico, A., Anikin, A., Romanenko, R. (2016). An ontological user model for automated generation of adaptive interface for users with special needs. In *2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA)* (pp. 1–6): IEEE.

Lafferty, J., McCallum, A., Pereira, F.C.N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Liu, J., Shahroudy, A., Xu, D., Kot, A.C., Wang, G. (2017). Skeleton-based action recognition using spatio-temporal lstm network with trust gates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(12), 3007–3021.

Luvizon, D.C., Tabia, H., Picard, D. (2017). Learning features combination for human action recognition from skeleton sequences. *Pattern Recognition Letters*, *99*, 13–20.

Ly, K.H., Ly, A.-M., Andersson, G. (2017). A fully automated conversational agent for promoting mental well-being: a pilot rct using mixed methods. *Internet Interventions*, *10*, 39–46.

Mavropoulos, T., Meditskos, G., Kamateri, E., Symeonidis, S., Tzimikas, D., Papageorgiou, L., Eleftheriadis, C., Adamopoulos, G., Vrochidis, S., Kompatsiaris, I. (2019). A smart dialogue-competent monitoring framework supporting people in rehabilitation. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments* (pp. 499–508): ACM.

Metz, C.E. (2008). Roc analysis in medical imaging: a tutorial review of the literature. *Radiological Physics and Technology*, *1*(1), 2–12.

Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781.

Münzner, S., Schmidt, P., Reiss, A., Hanselmann, M., Stiefelhagen, R., Dürichen, R. (2017). Cnn-based sensor fusion techniques for multimodal human activity recognition. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers* (pp. 158–165): ACM.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*(1), 3–26.

Nweke, H.F., Teh, Y.W., Mujtaba, G., Al-Garadi, M.A. (2019). Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions. *Information Fusion*, *46*, 147–170.

Pennington, J., Socher, R., Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv:1802.05365.

Pragst, L., Miehle, J., Minker, W., Ultes, S. (2017). Challenges for adaptive dialogue management in the kristina project. In *Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents* (pp. 11–14): ACM.

Ravindranath, P.A., Hong, P., Rafii, M.S., Aisen, P.S., Jimenez-Maggiora, G. (2018). A step forward in integrating healthcare and voice-enabled technology: Concept demonstration with deployment of automatic medical coding model as an amazon "alexa" skill. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *14*(7), P955.

Rhif, M., Wannous, H., Farah, I.R. (2018). Action recognition from 3d skeleton sequences using deep networks on lie group features. In *2018 24th International Conference on Pattern Recognition (ICPR)* (pp. 3427–3432): IEEE.

Richman, L.S., Kubzansky, L., Maselko, J., Kawachi, I., Choo, P., Bauer, M. (2005). Positive emotion and health: going beyond the negative. *Health Psychology*, *24*(4), 422.

Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J. (2013). Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, *105*(3), 222–245.

Sanderson, R., Ciccarese, P., Young, B. (2017). Web annotation data model.

Sang, E.F., & De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. arXiv:0306050.

Savino, J.A., & Latifi, R. (2019). Hospital and healthcare transformation over last few decades. In *The Modern Hospital* (pp. 23–29): Springer.

Stisen, A., Blunck, H., Bhattacharya, S., Prentow, T.S., Kjærgaard, M.B., Dey, A., Sonne, T., Jensen, M.M. (2015). Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems* (pp. 127–140): ACM.

Tai, L.K., Setyonugroho, W., Chen, A.L. (2020). Finding discriminatory features from electronic health records for depression prediction. *Journal of Intelligent Information Systems*, *55*(2), 371–396.

Tanaka, H., Adachi, H., Ukita, N., Ikeda, M., Kazui, H., Kudo, T., Nakamura, S. (2017). Detecting dementia through interactive computer avatars. *IEEE Journal of Translational Engineering in Health and Medicine*, *5*, 1–11.

Tanaka, H., Negoro, H., Iwasaka, H., Nakamura, S. (2017). Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. *PloS one*, *12*(8), e0182151.

Tang, Y., Tian, Y., Lu, J., Li, P., Zhou, J. (2018). Deep progressive reinforcement learning for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5323–5332).

Tran, T.N.T., Felfernig, A., Trattner, C., Holzinger, A. (2020). Recommender systems in the healthcare domain: state-of-the-art and research issues. *Journal of Intelligent Information Systems*,1–31.

Tsanousa, A., Chatzimichail, A., Meditskos, G., Vrochidis, S., Kompatsiaris, I. (2020). Model-based and class-based fusion of multisensor data. In *International Conference on Multimedia modeling, 5-8 January 2020*: Springer.

Tsanousa, A., Meditskos, G., Vrochidis, S., Kompatsiaris, I. (2019). A weighted late fusion framework for recognizing human activity from wearable sensors. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)* (pp. 1–8): IEEE.

Ultes, S., & Minker, W. (2014). Managing adaptive spoken dialogue for intelligent environments. *Journal of Ambient Intelligence and Smart Environments*, *6*(5), 523–539.

Vemulapalli, R., Arrate, F., Chellappa, R. (2014). Human action recognition by representing 3d skeletons as points in a lie group. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 588–595).

Wang, J., Liu, Z., Wu, Y., Yuan, J. (2012). Mining actionlet ensemble for action recognition with depth cameras. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1290–1297): IEEE.

Xia, L., Chen, C.-C., Aggarwal, J.K. (2012). View invariant human action recognition using histograms of 3d joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (pp. 20–27): IEEE.

Yu, Z., Black, A.W., Rudnicky, A.I. (2017). Learning conversational systems that interleave task and non-task content. arXiv:1703.00099.

Yu, Z., Xu, Z., Black, A.W., Rudnicky, A. (2016). Strategy and policy learning for non-task-oriented conversational systems. In *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue* (pp. 404–412).

Zanfir, M., Leordeanu, M., Sminchisescu, C. (2013). The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2752–2759).