Review Article

iMedPub Journals
http://www.imedpub.com

American Journal of Computer Science and Information Technology

2021
Vol. 9 No. 10: 113

Voice User Interface: Literature Review, Challenges and Future Directions

Rakotomalala Francis[1*], Hajalalaina Aimé Richard[1], Ravonimanantsoa Ndaohialy Manda Vy[2], Randriatsarafara Hasindraibe Niriarijaona[1]

[1]Laboratory for Mathematical and Computer Applied to the Development Systems (LIMAD), University of Fianarantsoa, Fianarantsoa, Madagascar

[2]Department of Engineering and Innovation Innovation Sciences and Techniques (STII), University of Antananarivo, Antananarivo, Madagascar Sciences and Techniques (STII), University of Antananarivo, Antananarivo, Madagascar

Corresponding author:
Rakotomalala Francis, Laboratory for Mathematical and Computer Applied to the Development Systems (LIMAD), University of Fianarantsoa, Fianarantsoa, Madagascar

francis_rakotomalala@ymail.com

Citation: Francis R, Richard HA, Manda RN, Niriarijaona RH (2021) Voice User Interface: Literature Review, Challenges and Future Directions. Am J Compt Sci Inform Technol Vol.9 No.10: 113.

Abstract

Natural user interfaces are increasingly popular these days. One of the most common of these user interfaces today are voice-activated interfaces, in particular intelligent voice assistants such as Google Assistant, Alexa, Cortana and Siri. However, the results show that although there are many services available, there is still a lot to be done to improve the usability of these systems. Speech recognition, contextual understanding and human interaction are the issues that are not yet solved in this field.

In this context, this research paper focuses on the state of the art and knowledge of work on intelligent voice interfaces, challenges and issues related to this field, in particular on interaction quality, usability, security and, the presentation of new emerging technologies in this field will be the subject of a section in this work.

The main contributions of this paper are overview of existing research, analysis and exploration of the field of intelligent voice assistant systems, with details at the component level, identification of areas that require further research and development, with the aim of increasing its use, various proposals for research directions and orientations for future work, and finally and study of the feasibility of designing a new type of voice assistant and general presentation of the latter, whose realization will be the subject of a thesis.

Keywords: Voice assistant; natural language processing; artificial intelligence; machine learning, human-machine interaction; review.

Received: October 14, 2021; **Accepted:** October 28, 2021; **Published:** November 04, 2021

Introduction

People have already witnessed a method of interaction between a user and mobile devices, where buttons played an important role in the interaction with the device. However, this interaction has been developed from the past in different ways, as shown by the evolution of smartphones, tablets and smart devices. Buttons have been replaced by a touch screen system, as well as many new methods of interaction with a mobile device, such as voice, camera and fingerprints.

The focus is now on new ways of communicating so that man and machine can work together without apprehension. Intuitive, simple and recognizable are the terms used to describe human-machine interfaces. Researchers are working to recognize and interpret gestures, facial expressions and emotions. Even brain waves are recorded by neuron-sensors in order to interpret the intentions of individuals. Scientists are even working on the spoken language between man and machine. Based on the model of smartphone voice assistants, the machines will be able to react to spoken instructions.

However, despite the fact that much progress has been made since the first human-machine interfaces were created, interaction is still limited by the autonomy of machines. Human-machine interaction is therefore still far from achieving what is called “natural interaction”.

A natural user interface (NUI) is a human- computer interaction system that the user uses through intuitive "invisible" actions.

The purpose of these interfaces is to hide the complexity of the system even if the user is experienced or the interactions are complex. Examples of actions commonly used by NUI include touch and gestures. In recent years, a new generation of voice-activated personal assistants has become common and widespread. Indeed, communication with devices using voice is now a common task for many people. Intelligent personal assistants, such as Amazon Alexa, Microsoft Cortina, Google Assistant, or Apple Siri, allow people to search for various topics, schedule a meeting, or make a call from their car or their hands, no longer needing to contain mobile devices. These intelligent

assistants therefore use natural language user interfaces (NLUI) to interact with users. NLUI involves the translation of human intention into commands to control devices via voice recognition. These gadgets feature advances in artificial intelligence (AI), speech recognition, semantic web, dialogue systems and natural language processing, thus consolidating the concept of intelligent assistant. The term "intelligent voice assistant" therefore translates as a system that is able to understand, respond to voice input and process the user's request. As people interact with an increasing number of devices by voice, conversation becomes an essential mode of interaction between humans and computers. The benefits are not only the voice control, but also the dialogue-style nature of the interactions, which allows hands-free human-device interaction, and this technology allows computing to work in new and unexplored areas.

Since the introduction of voice assistant systems in the 1990s, many research papers have been published on the subject, including some surveys. In [1], the authors present an overview of intelligent voice assistant types, devices and user interface, indicating their current challenges. Pokojski [2] describes the concepts of voice assistant software, including the proposal of a specific approach and the discussion of knowledge-based systems. The authors conclude that it is relatively difficult to develop universal intelligent personal assistant software. Another paper focuses on comparing personal assistants to understand the design, architecture and implementations of the framework [3]. Finally, [4] explores cognitive assistants, which is a subset of intelligent assistants, focusing on pervasive platforms and services.

In analyzing all the articles found in this research area, few people have proposed systematic literature reviews aimed at identifying different key areas of research on intelligent voice assistants. Although work of this nature is essential in many areas, in addition to artificial intelligence, expert systems, conversational and cognitive agents, as it identifies the main applications, technologies involved software architectures, challenges and open questions, as well as opportunities regarding the voice assistant. Therefore, one recently published work [5] aims to discuss important concepts and conclusions regarding intelligent assistants, using this methodology. However, as noted in other work [6,7], systematic review literature searches have some limitations due to the enhanced methodological rigor. Furthermore, these works are limited to aspects related to concepts and applications. Shortly after, another research paper [8] focused on the knowledge of virtual environment and virtual assistant interfaces work, and presents virtual assistant applications that help to access the software without having knowledge of how the software works. It also describes the limitations and challenges that arise in virtual assistant technology.

In this work, they tried to mitigate the search bias in order to obtain an overview of the current literature without explicitly evaluating articles that refer exclusively to the improvement of the intelligent voice assistant components. As a result, the literature is still limited to aspects related only to the system rather than the inclusion of new techniques such as speech recognition or the form of interaction, as examples.

In this paper, we will provide an overview of the state of the art of intelligent voice assistants with the aim of analyzing; in addition to presenting the research directions and new technologies related to; the architectures and components of these assistants.

Obviously, the current study is therefore the result of a systematic review of the literature designed to provide an overview of the research field of intelligent voice assistants, emphasizing especially, unlike the last reviews, those components and architectures of current voice assistant systems; and identifying the different specific and general promising directions.

Given the state of the art studied, this paper not only highlights the challenges related to usability, security and privacy, but also proposes perspectives for solutions. The proposals made thus provide a set of contributions that can be used as guidelines for future research by the scientific community.

The work is organized as follows: Section 1 discusses the context of the intelligent voice assistant. Subsequently, Section 2 presents the results obtained from the collection of articles by highlighting the current literature of the different technical components of the intelligent voice assistant. New technologies in the field are presented in Section 3. In section 4, we will highlight recent challenges before identifying future research directions and orientations that can bring improvement on the intelligent voice assistant system. In the last section, we will make a general description of our main future work which is nothing but the result of the ideas from related work, perspectives and limitations of the research work of different recently published papers.

Context of the intelligent voice assistant

Human-computer interaction 1 by voice is not a new idea. For a human being, speech is the most natural means of communication, so the idea of teaching computers to recognize and understand verbal language may be as old as the first computers. However, early attempts to do so were limited by the computing power and storage capacity of computers, so that computers could only recognize a small subset of words [9]. Modern AI assistants can have a fluid conversation with users, providing the same experience as if users were talking to a real person.

Definition: In a nutshell, a voice assistant can be described as software that receives and interprets verbal input from the user, performs a requested task if necessary, and then responds back to the user in a verbal format. Voice assistants rely on artificial intelligence and self-enhancing machine learning algorithms to improve the accuracy of speech understanding. Virtual assistants can be distributed as a stand-alone device (smart speaker) or as an application for smartphones or computers.

Architectures standards: The relevance of a robust architecture for intelligent assistants is in fashion, indeed, several proposals for architectures have emerged in recent years [10-12]. These authors draw attention to the need for skill- extensible assistant behaviors to provide both flexibility and scalability. Thus,

personal assistant behaviors actually use the concept of cloud computing, which is enabled by web services distributed over many repositories with application services provided by different companies, organizations or developers.

As in other work [13], the main agents of intelligent assistants use the process described in **Figure 1**. These agents have as a first step the transcription of speech into words or sentences to allow the use of natural language understanding and resources, leading to a semantic interpretation of the input.

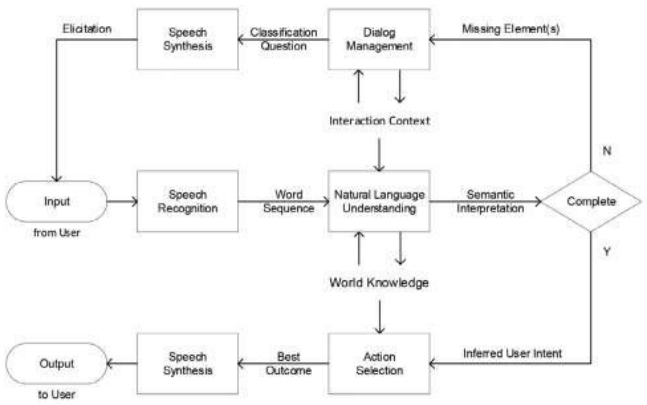

Figure 1 An overview of the underlying intelligent voice assistant interaction model.

Some work proposed advances in personal assistant architectures [14–16], frameworks [17–19], advances in techniques for predicting user behavior [20], approaches to better tracking intentions [21] and extending the assistant to other languages [22]. In addition, some papers proposed techniques to recommend the right information at the right time and help intelligent personal assistants to perform tasks [23], to predict the user's next intention and to make the conversation friendlier. In sequence, Milhorat et al [24] proposed a set of challenges for PDAs that could improve their state-of-the-art context- aware architecture by using the large amount of data available and not only according to what is directly requested by the user. These challenges have already been overcome by some commercial personal assistant applications, including Google Assistant, but they pose a user privacy problem due to the need to store a lot of information such as real- time location, speech, etc. [25].

The majority of the work surveyed to propose architectures used an already known platform called Sirius, which is "an open end-to- end intelligent assistant web service application that accepts requests in the form of voice and images, and responds with natural language". Sirius consists of three services: automatic speech recognition, question answering and image matching. The first uses a combination of a Hidden Markov Model and a Deep Neural Network (DNN) [26]. In the second service, the text output of the automatic speech recognition is used as input to Open Ephedra [27]. Finally, they based the image matching service on the SURF algorithm [28].

Components of the voice assistant architecture

Voice recognition: Everything we say, whether verbally or in writing, contains huge amounts of information. In theory, we can understand and even predict human behavior using this information. The problem is that a person can generate hundreds or thousands of words in a statement, each sentence having its corresponding complexity.

Thus, we need a comprehensive system to structure and analyze this data to extract relevant information, to be used for various applications, reducing manual work.

Speech recognition is an interdisciplinary subfield of computational linguistics that develops methodologies and technologies for the recognition and translation of spoken language into text by computers.

For many decades, researchers have been working in the field of speech recognition and communication. Researchers have made valuable contributions in this field so far. They have also contributed in the field of speech recognition as well as speaker-dependent and speaker-independent speech recognition.

In this respect, a work [29] presents the main scientific techniques and approaches, which have been developed by various researchers in the broad field of speech recognition, that have developed over the last decades. The fundamental techniques and methods for speech recognition are discussed.

Speech recognition: Speech recognition technology can be used to perform an action based on human-defined instructions. The human needs to train the speech recognition system by storing the speech patterns and vocabulary of their language in the system. In doing so, they can essentially train the system to understand them when they speak. It incorporates knowledge and research from the fields of linguistics, computer science and electrical engineering. The five phases of NLP2 involve lexical analysis (structure), syntactic analysis, semantic analysis, discourse integration and pragmatic analysis [30].

To shed more light on this problem, an article [31] distinguished between a review of the components of NLP, followed by a presentation of the history and methods of NLP, and the state of the commercial prospects of natural language processing for speech recognition.

Indeed, these many years of research have made automatic speech recognition one of the challenging areas and have made it an important research topic.

In another work [32], the authors present some of the important work or updates done by researchers to make this automated speech recognition (ASR) work as it is today.

The literature survey shows the chronic order of inventions and discoveries in the field of speech recognition.

Recently, deep learning algorithms have mainly been used to further improve the capabilities of computers to understand what humans can do, which includes speech recognition. Therefore, it is quite natural that one of the first applications of deep learning was speech, and to date, a large number of research papers have been published on the use of deep learning for speech-related

applications, especially speech recognition [33–35].

Therefore, the paper [36] provides an in-depth review of the various studies that have been conducted since 2006, when deep learning first emerged as a new area of machine learning for speech applications.

Among other things, nowadays all kinds of devices in life are oriented towards networking and artificial intelligence, and the mode of operation has also shifted from touch buttons to more advanced control methods such as voice control.

Therefore, in terms of the requirements of intelligent speech recognition in the family, an acoustic model based on speech feature recognition and the adopted DNN-HMM is designed in [37]. After experimentation, the speech recognition accuracy of the speech recognition system model reached the design requirements.

Although machines currently share the same level of ASR cognition with humans in quiet environments, they do not yet share the same level of ASR cognition with humans in noisy work spaces [38]. By simulating the mechanism of the human brain on multimodality-based speech recognition, audiovisual speech recognition (AVSR) has been put forward.

Therefore, a detailed review of recent advances in the field of audiovisual speech recognition is presented in [39]. Robust AVSR has been shown to significantly improve the recognition accuracy of ASR systems under various adverse acoustic conditions [40].

Speaker recognition: On the other hand, speaker recognition has also been the subject of active research for many years and has various potential applications such as in mobile phones and many biometric security systems. The speaker recognition process aims at extracting the identity of the speaker based on individual information from the speech signals.

In this respect, a paper [41] provides a detailed overview of the different machine learning algorithms used for the speaker identification task.

These ML3 algorithms include some conventional approaches. Secondly, the most widely used methods nowadays using deep learning such as deep neural network (DNN), deep belief network (DBN), convolutional neural network (CNN). Among all ML techniques, deep learning has achieved a state-of-the-art result in the field of speaker identification.

Moreover, it is clear that it is a difficult task to teach a machine the differences of human voices, especially when people belong to different backgrounds such as gender, language, accent, etc.

Therefore, a paper [42] recently used the deep learning approach to build and train two models: Artificial Neural Network (ANN), Convolutional Neural Network (CNN).

In the former, the neural networks are fed with various features extracted from the audio collection, while the latter is trained on spectrograms. Finally, the transfer learning approach is deployed on both to obtain a viable result using limited data.

Furthermore, with the increasing demand for human-computer interaction (HCI) systems, speech processing plays an important role in improving HCI systems. The development of gender recognition systems has applications in gender-based virtual assistants, telephone surveys and voice-activated automation systems.

Therefore, another project [43] used the analysis of speech signals and its different parameters to design a gender classifier. The main objective of the gender classifier is to predict the gender of the speaker by analyzing different parameters extracted from the speech sample.

Finally, while some limited research has been conducted in the broad research area of speaker recognition, most of it is currently outdated. Therefore, an article [44] focuses on the renowned research area and explores various dimensions of such an exciting research field. The paper targets the research area from several aspects, including fundamentals, feature extraction methodologies, datasets, architectural constructs, performance measures and challenges.

Research Directorates: As seen above, there is a lot of work trying to solve the automatic voice recognition problem, but according to, the majority of these papers have focused on speech recognition as an application in use.

The analysis of the results provided in, allowed us to identify new and interesting research topics that have not yet been examined, as well as to highlight some of the gaps in the existing studies.

Firstly, it is worth trying, when using deep learning models, other feature extraction methods such as linear predictive coding (LPC) 4.

Furthermore, 75% of the DNN models were stand-alone models while only 25% of the models used hybrid models. It would therefore be more appropriate to use hybrid models as research has shown that using Hidden Markov model (HMM) or Gaussian Mixture Modeling (GMM) to inform a DNN model gives better results [45].

Another observation made in this study too is that there is little work on speech recognition using recurrent neural networks (RNN). There is great interest in conducting research using deep RNN in the future, as RNN models, especially long and short term memory (LSTM), are very powerful in speech recognition [46].

We therefore plan to employ recent types of RNNs in speech recognition tasks to train efficient and robust speech recognition models.

Furthermore in it was shown that robust audiovisual speech recognition significantly improves the recognition accuracy of speech recognition systems under various adverse acoustic conditions.

The analysis also showed that deep learning and AVSR are now inextricably linked, and the favorable anti-noise performance of the end-to-end AVSR model and the deep-level feature extraction capabilities of deep learning-based feature extractors will be able to guide a class of multimodal HMIs directly to a solution.

As part of the continuity of the field, we will focus on the exploitation of deep learning techniques for speech recognition. Then, we propose to introduce other analysis variables to widen the scope of recognition in natural interfaces.

In addition, we plan to study the feasibility of multilingual speech recognition to overcome the limitation of speech restriction.

However, we have seen earlier that speech recognition is the emerging area of security and authentication for the future.

As such it provides users with the appropriate and effective method of the authentication system based on voice recognition. Therefore, an interesting direction would be to secure the use of the virtual assistant from the voice interface itself. In addition, we plan to introduce also the gender modality of the user in the processing and analysis of speech.

Indeed, voice is used in the fields of human voice authentication for security purposes and identification of a person among a group of speakers.

With this in mind, several papers have studied different machine learning algorithms used for the speaker identification task.

It is observed that unlike the complicated process of extracting features from speech signals and then feeding these features into early classifier models such as SVM, RF and kNN, the SI process feeds the raw audio signals directly to DL-based classifiers such as CNN.

This has led to good tangible results in terms of performance and efficiency. Not surprisingly, with the continuous progress of AI, the speaker identification task should become more accurate and less complicated. Although we still propose to identify the best context-dependent ML algorithms for a robust speaker recognition system based on the problem.

Finally, recently, significantly confirms the existing defaults and variants of ASR systems that could help us to quickly adapt the concepts of the research area.

Furthermore, comparisons and future guidelines would help to explore a broader perception of speaker recognition technologies, including architectural and feature extraction terminologies.

Finally, we propose to include speaker analysis in speech recognition to target messages and speakers.

Natural language understanding

The way the machine analyses natural language is not quite the same as that of a human. Indeed, as the amount of speech or text to be analyzed increases, the processing becomes more complex. Several methods for automatic natural language understanding have been proposed over the years. Recently, researchers have focused on lexical and semantic relationships through the extraction of keywords or, more recently, key phrases in order to deduce the overall significance of a text, a paragraph or a sentence. They can even go further by generating summaries of the latter by exploiting the keywords or phrases obtained.

Keyword extraction: Keywords, which are important expressions in documents, play an important role in various applications of text mining, information retrieval (IR) and natural language processing (NLP). The task of keyword extraction is an important problem in text mining, IR and natural language processing.

Although several works on keyword extraction have been published in the last few years, since our paper is about voice assistant we have done some sorting and will present only those papers that we consider relevant to this topic.

To date, according to a work [47] which provides a general and comprehensive introduction to the field of keyword/key phrase extraction, no single method or set of features can yet efficiently extract keywords in all different applications.

Therefore, we will present some recent state- of-the-art methods, some using supervised approaches and others using unsupervised approaches.

An interesting work relevant to our application area dates from 2015 [48] where the authors built a corpus of Twitter tweets annotated with keywords using crowdsourcing methods, which was then used to build and evaluate a system to automatically extract keywords from Twitter.

Indeed, these messages tend to be shorter than web pages, where content must be limited to 140 characters. The language is also more informal with many messages containing spelling errors, slang, abbreviations among domain-specific artifacts. In many applications, existing datasets and models tend to perform significantly worse on these domains. The basic system, defined using existing methods applied to the dataset, is significantly improved using unsupervised feature extraction methods.

Technically, there are still many difficulties in text categorization in a high dimensional feature space [49]. When whole words in the document have been used as training features, the computational complexity will be considerably increased, making the text categorization task transformed into a computationally intensive type of task [50].

As a result, a study [51] designed a supervised keyword classification method based on the Text Rank 5 automatic keyword extraction technology and optimizes the model with the genetic algorithm to contribute to the text classification to model the topic functionality. This method is interesting in that it is a guiding and practical meaning for keyword and text classification based on keyword characteristics.

Like the methods outlined above, the aim of supervised approaches is to train a classifier on documents annotated with keywords to determine whether a candidate word is a keyword or not. As the interaction between the user and the virtual assistant is quite diverse and dynamic, it would be more appropriate to study unsupervised approaches to keyword extraction.

In general, automatic keyword extraction techniques can be broadly classified into two classes: the Vector Space Model (VSM) and the Graph-Based Model (GBM) [52]. VSM algorithms are quite efficient for keyword extraction, but do not focus on the class label information of the classified data.

Various unsupervised approaches focus on graph-based ranking methods for keyword extraction. However, the main limitation of these approaches is that they focus more on the single score of a word, i.e. the frequency of occurrence of a word in the document and less on the impact and position of the word in a sentence.

Word frequency analysis is another popular method of keyword extraction proposed by Jones [53]. A term with a higher TF- IDF 6 score indicates that it is both important to the document and relatively rare in the corpus. However, a major drawback of this method is that if the frequency of use of the term is low in a particular document, then this term will have a low TF, which reduces its possibility to be selected as a keyword.

Therefore, a work [54] extended existing state-of-the-art methods with additional features to capture the frequency and relative position of words. This work uses fuzzy logic to extract keywords from individual documents. Each word in the document is assigned a weighting of its position in particular sentences throughout the text document related to its frequency of occurrence. Words with a higher weight will be selected as candidate keywords for the document.

However, another work [55] attempts to explore the advantages of the VSM and GBM techniques by proposing a keyword extraction using the KESCT (Supervised Cumulative Text Rank) technique. The paper exploits a new statistical term weighting mechanism, called Unique Statistical Supervised Weight (USSW), and incorporates it into Text Rank, replacing the TF-IDF algorithm.

Key phrase extraction: In a real natural interaction, at a certain point it happens that taking into account only words as analysis factors is no longer sufficient. Therefore, to boost the information to be exploited, an agent capable of extracting key phrases is a plus for a complete voice assistant with more flexibility. We therefore consider it necessary to review recent state-of-the-art methods for key phrase extraction.

To begin with, there are a number of noteworthy key phrase extraction surveys. For example, [56] focus on the mistakes made by state-of-the-art key phrase extractors. Although their analysis is not based on a large number of documents, it is quite interesting and well presented.

An interesting paper [57] introduces key phrase extraction, provides a well-structured review of existing work, offers interesting insights into different evaluation approaches, highlights open issues and presents a comparative experimental study of popular unsupervised techniques on five data sets.

An unsupervised AKE 7 overcomes the critical challenges of corpus formation and domain bias by making the retrieval task a ranking problem.

To shed more light on the problem of key phrase extraction, the main unsupervised methods for extracting key phrases are summarized in [58], and the authors analyze in detail the reasons for the differences in the performance of the methods, and then they have provided solutions.

Thus to help researchers further improve the performance of the method in the key phrase extraction task, they introduce new features that may be useful.

Compared to other NLP tasks, unsupervised approaches to perform AKE have difficulty in achieving a better result [56], due to the complexity of AKE tasks, which require not only local statistical information about the terms contained in the document, but also background knowledge to capture the relationships between them [59]. Many recent approaches suggest using external knowledge sources, such as WorldNet [60] to obtain rich information about relationships during AKE [61,62]. Although these approaches demonstrate improved AKE performance in some cases, their knowledge sources used are not sufficiently consistent to provide general information about terms in an arbitrary domain, and therefore a term representative of the document's topic may be ignored simply because the knowledge source does not store information about it.

Therefore, the authors of the paper [63] presented same cluster, an unsupervised clustering-based key phrase extraction method that solves the coverage limitation problem by using an extensible approach that integrates an internal ontology (i.e., WorldNet) with other knowledge sources to acquire a broader knowledge base.

The results not only show that same cluster outperforms the compared methods but also verify Liu's [64] conclusion that AKE methods based on unsupervised clustering can be efficient and robust even in several domains.

In general, unsupervised statistical techniques are computationally expensive due to their large number of complex operations, and unsupervised graph-based techniques perform poorly due to their inability to identify cohesion between words that form a key phrase.

For this reason, a new unsupervised automatic key phrase extraction technique, called Tree- based Key Phrase Extraction Technique (TeKET) [65] has been proposed, which is domain and language independent, uses limited statistical knowledge, but no train data is required.

In addition, there have been significant advances in deep contextual language models [66-67]. These models can take input text and provide contextual embedding for each token for use in downstream architectures. They have been shown to achieve state-of- the-art results for many different NLP tasks. More recent works [68-69] have shown that contextual embedding models trained on domain-specific corpora can outperform general purpose models.

Despite all the developments, there has been no work on the use of contextual plunging for key phrase extraction.

To explore the hypothesis that key phrase extraction can benefit from contextual embedding, [70] approach key phrase extraction as a sequence labeling task solved using a BiLSTM-CRF [71], where the underlying words are represented using various contextual embedding architectures. The authors demonstrate that contextual embedding significantly out perform their fixed counterparts in key phrase extraction.

Indeed, compared with hand-designed features and traditional discrete feature representation, the deep learning technique offers a different way to automatically learn the representation of dense features for text, such as words, and sentences.

Recently, the sequence-to-sequence based generative framework (Seq2Seq) is widely used in the key phrase extraction task, and it has achieved competitive performance on various benchmarks. The main challenges of Seq2Seq methods lie in acquiring an informative representation of the latent documents and in better modeling the compositionality of the target key phrase set, which will directly affect the quality of the generated key phrases.

Therefore, the authors of the work [72] propose to adopt dynamic graph convolutional networks (DGCN) to solve the above two problems simultaneously.

Text synthesis: The volume of text retrieved by the speech interface can increase as the user needs it until it becomes large or at some point requires a lot of effort and time to go through, so by providing a summary, it reduces the reading time and only the significant points are highlighted in the content.

Two major techniques have been proposed for computing text summarization. First,extractive text summarization, which focuses on extracting key phrases from the data to create a summary without changing the original text. Second, abstract text summarization, which focuses on understanding the appropriate meaning of given natural sentences and generating a natural language.

As in the previous sections, we will limit ourselves to an overview of works that are more or less relevant to our research area.

Let's take a similar problem of using the Chabot to manage conversations. The Chabot

8 can answer a large volume of frequent and typical questions, which will certainly reduce the company's human resource consumption. However, sometimes a Chabot cannot satisfy users with the answers provided; in such a situation the Chabot should be transferred to an agent to handle the rest of the conversation. A concise summary of the user's statements helps the agent to handle the rest of the discussion without causing annoying delays. Since the dialogue between Chabot and the user is fragmented and informal compared to the classical summarization dataset, models of high complexity often struggle to achieve ideal results.

With this in mind, the authors in [73] present an extractive chat summary system to provide a concise summary of the topics discussed in the chat. They used supervised and unsupervised methods for keyword extraction, combinations of features with fine tuning, and Red metrics to compare the generated summary with the reference summary.

Later, another paper [74] presents various techniques for text summarization and classification by extraction.

The advantage of this approach is that it gives better results as it focuses on the key phrases to be extracted rather than on generating natural language by understanding the appropriate meaning of the data.

Extractive text summarization also provides an advantage to the system by offering better performance in terms of fluency and grammar.

However, extractive text summarization lacks redundancy within sentences and consistency between sentences. Furthermore, the accuracy of the extraction depends on the chosen affinity function. Although,this accuracy can be improved in the future by using different affinity functions.

Research Directorates: Since the technique proposed in is domain and language independent. In future work, it could be used to perform many NLP tasks such as document summarization. We plan to apply the extraction method to extract keywords in an interaction.

After analyzing the work [75], the results can be used to create a convenient basis for quick keyword searches of trials. It is also possible to generate a summary description to reconstruct the exact theme of each group from extracted keywords.

In the future, the study should be pursued in the direction of identifying the correlation between clustering constituents and lexical diversity values, i.e. towards a greater focus on the search task of automated evaluation of lexical features in text.

In the context of the voice assistant, it is interesting to apply this approach to summarize the interaction between human and machine. In the field of IT entity management, we can also use the approach to perform a fast extraction of keywords related to the IT field.

Then, the work can be extended by considering the formatting present in the text such as bold, italics, underlines, and font size and font style. A weighting can also be assigned to the labels of the images in the text, which could lead to more refined and flexible results.

More work could be done to study the method on different languages and scripts. One idea is to use the approach to study the Malagasy language if one wanted to create voice assistants accepting the Malagasy language.

Afterwards, we have seen that the KESCT model fills the various gaps of the VSM and GBM keyword extraction techniques.

In the future, the KESCT model is bound to have a wide scope of extension and application to different case studies and real time applications. This is very important for problems that are solved in real time such as the case of virtual assistants. We therefore focus on adapting this model for real-time analysis of interactions with virtual assistants.

On the other hand, the approach [76] allows efficient keyword search using n-gram string similarity algorithms, even when the keywords are subdivided into several words in the corpus text.

This technique is interesting in that it allows the extraction of fuzzy keywords in an extreme environment such as natural interactions like conversations.

Similarly, there is a lot of work on key phrase extraction that shows the need for improvement but also several interesting

leads in the field of virtual assistant.

First, the technique proposed in is domain and language independent. On the other hand, although Sem Cluster performs better than other approaches, there is still room for improvement.

WorldNet should be extended with more personalized knowledge sources and their impact on performance should be investigated using personal documents with greater length and domain variance (emails, health records, micro blogs) than the currently used datasets.

For our part, we propose to introduce the WSD 9 problem to find word equivalences in the digital management domain.

Similarly, the extraction method proposed in is also domain and language independent. This leads us to propose to use the unsupervised approach for the extraction of informal keywords or phrases (e.g. words or phrases from combinations of two different languages.

Furthermore, the results of the review show that simple unsupervised methods are solid baselines that should be considered in empirical studies and that deep learning methods provide state-of-the-art results. We especially want to cross the frontier of deep learning and unsupervised language models [78] since the exploitation of such models for key phrase extraction and/or generation appears to be the most interesting future direction.

As reported in [79], the application of the self- training method of the deep learning model has further improved the performance and achieved competitive performance with previous state-of-the-art systems.

We therefore plan to use self-learning with the interaction data to increase the performance of the extraction model.

Still in the context of deep learning, the model proposed in, improved by a diversified inference process, can generate even more accurate and diversified outputs. We therefore believe that the application of this approach in the tasks of voice assistant systems is relevant to improve the understanding of voice interactions.

Another approach that has given interesting results is presented in [80]. This approach can be easily transferred to other contexts and languages: the use of pre-trained models allows the use even on small data sets, the clustering algorithms are standard and easily adaptable.

The pipeline identified for the Italian language allows easy adaptation to other languages such as French, German, etc. However, mainly for Italian, especially in the case of specific topics such as recipes, the pre- trained models struggle to be adequate and to provide correct results.

According to the authors, the aspects that will be further investigated and analyzed in the future concern: (1) testing the approach on French, German texts, evaluating the results obtained and the effort needed to adapt it, (2) testing FastText10 and other word embedding models, (3) adapting and customizing the pre- formed embedding models with the inclusion of new terms, (4) improve/integrate the evaluation methods, able to take into account the different aspects of the problem, (5) integrate Wikipedia, the largest human collected and organized encyclopedia on the Web, as a knowledge base to measure the relationship between terms.

With respect to our virtual assistant problem, it would be interesting to adopt the approach

[80] to deal with multilingual or even linguistically hybrid interactions or conversations.

Then, we plan to test the approach on Malagasy texts. Another significant direction is also to adapt and customize the pre-trained integration models with the inclusion of new terms, to add out-of-vocabulary words and update weights, in order to solve the problem of specialized words or contexts. These can be informal terms or recently used terms, or new word meanings.

Furthermore, according to [58] one of the common problems in the evaluation of retrieval methods is the impact of the gold standard. In order to reduce it, it is necessary in the future to introduce external knowledge to determine whether the key phrases extracted by a method and the gold standard key phrase have the same semantics.

On our side, we propose to introduce the notion of equivalence of expressions and sentences according to the interaction domain.

After the extraction of key words/phrases, we still need a good automatic summarization system to complete the natural language understanding agent of our voice assistant. Earlier, we presented some interesting articles on the subject and after some analysis we could extract some guidelines for future work.

First, the extracted keywords are usually terminology for more technical domains [81], such as the case of intelligent assistants. Human experts are still needed to illustrate the additional meaning of these raw keywords.

This problem could be modified by implementing the algorithm on several interactions in the same domain and analyzing the extracted keywords together.

Moreover, there is nothing to prevent the method presented in from being applied only in the scientific domain. It will be interesting to extend the method to more general domains and to find indications for it. For example, we can apply this method to target discussions that may contain information about possible system commands to be made from a long conversation.

We also propose to use algorithms of lower complexity to save time and material for the synthesis of discussions or interactions. Moreover, for summary generation, these often give better results than complex summary generation methods. However, different algorithms have different optimization methods and the quality of the features will have a very critical impact on the effect of the learning process [73]. Therefore, we should take these constraints into account.

On the other hand, the results in showed that there are

fluctuations in the accuracy of determining the conversation as false due to misinterpretation of natural language resulting in a bias.

And this is the main advantage of this approach as it focuses on the key phrases to be extracted rather than on generating natural language by understanding the appropriate meaning of the data. This leads us to consider building on the approach by working only on sentences relevant to the management of digital entities. In other words, it makes more sense to neglect interactions or conversations that have no domain equivalence.

Conversational agent

State of the art: All assistance robots and Chabot, including those running Caresses-Cloud11, services, still have many limitations. One of the main reasons may lie in the efficiency of the algorithm responsible for choosing the topic of conversation according to the user's sentence: if the algorithm is not able to understand the context correctly, the agent's response will not be appropriate even if the system had the knowledge to respond consistently.

Therefore, a work [82] aims to improve the capabilities of knowledge-based conversation systems, making them more natural and enjoyable. This consists of developing a set of algorithms to improve the verbal capabilities of the CARESSES system.

In this work, the authors adopt different methods such as the brute force method, the method that exploits the entity type of the extracted concept, the method based on the definitions of the concept given by the user, and the method based on the category of a sentence concerning this concept. In addition, other tools such as the Dialog flow Web Service, Cloud Natural Language and Small Talk are also used. The results of this experiment showed that the changes introduced in the dialogue algorithm made the system much more consistent, more accurate, friendlier and less annoying.

Later, authors study the problem of imposing conversational goals/keywords on open- domain conversational agents, where the agent has to lead the conversation to a target keyword in a smooth and fast way.

For the first time in this task, [83] proposes to use CKG (Content Knowledge Graph) for keyword transition and then proposes two GNN-based models to incorporate common sense knowledge for next round keyword prediction and keyword augmented answer retrieval, respectively.

In this work, a large-scale open-domain conversation dataset for this task, obtained from Reedit 12, , is presented. The language models from Reedit are much more diverse than the ConvAI2 [84] used in existing studies, which are collected from hundreds of crowd workers.

Moreover, everything we express (verbally or in writing) contains huge amounts of information. In theory, we can understand and even predict human behavior using this information. However, a person can generate hundreds or thousands of words in a statement, each sentence having its corresponding complexity.

Thus, we need a comprehensive system to structure and analyses this data to extract the relevant information, to be used for various applications, reducing manual work.

To this end, a system [85] is proposed to recognize speech, process it to extract keywords, and analyses the accompanying sentiments.

Such systems help to improve user satisfaction, lead to fewer breakdowns in conversations and ensure user retention. Therefore, dialogue systems that are able to generate responses while taking into account the user's emotional state and feelings are the most desirable advancement in AI.

For this reason, another work [86] has the main objective of building a robust dialogue system or Conversational Artificial Intelligence (CAI) agent that should be able to communicate like a human. The authors aim to increase the robustness of the model with recent advances in artificial intelligence (AI), natural language processing (NLP) and machine learning (ML).

The Seq2Seq model [87] is the one dedicated to sequential data processing, which has a remarkable success in the machine translation task. And it shows great promise in other NLP problems such as text summarization, speech recognition, keyword extraction, etc. In which one of the main ones is the conversation system.

Following this path, in [88], the authors build a domain-specific generative Chabot using neural networks to train a conversational model that reads the data model and answers a new question. They show the application of the Seq2Seq model in Chabot systems.

Furthermore, in most PDAs13 currently in production, all assistant dialogues are still hard-coded rather than intelligently generated. The lack of diverse dialogues can make assistants rigid and unnatural, and it is tedious for development to maintain a large number of hard-coded dialogues.

Thus, a paper [89], inspired to generate PDA dialogues using paraphrasing, explores several deep learning architectures for the generic paraphrasing task.

In this work, the authors build on the original Seq2seq 14 + LSTM 15 model with bag-of- words (BOW) sampling proposed in [90], and experiment with variations of this model using the transformer architecture.

Research Directorates: In the tests of the proposed model that have been carried out provide information about the good ability of the system to implement an engaging, coherent and responsive verbal interaction, to recognize, and finally to acquire new concepts efficiently.

However, experience has shown that when using the developed dialogue algorithm, the average response time does not increase significantly after adding thousands of new concepts. In addition, further tests still need to be analyzed.

In our opinion, mixed (human-human) systems will also increase the quality of the machine's responses by drawing on human

responses in real time.

On the one hand, the dialog flow system can be used to capture messages in the design of the virtual assistant of the future. And on the other hand, the approach of adding new knowledge can be used in the generation of new commands to overcome the limitation of the actions of system assistants.

In addition, the introduction of informal language comprehension should be considered.

Secondly, personal and human evaluations show that the model presented in produces smoother responses and reaches the target keyword faster than competitive references. It seems to us therefore more appropriate to draw on this model in the prediction of system commands from human- machine interactions.

Furthermore, since the work on text processing in involves only unsupervised learning, these can be scaled, improved or modified according to different application requirements.

In the future, these models can be trained with datasets according to the need and application, and tested with various approaches to obtain the desired results for the development of a complete NLP system.

Therefore, we plan to use the proposed architecture in the design of a complete NLP system.

As mentioned above, recently, deep learning methods have had countless successes in the field of NLP. Indeed, in the proposed dialogue system has the gift of multimodal detection of feelings, emotions and sarcasm.

In another line of research, the authors focus on the development of multimodal dialogue agents for several applications, such as fashion, catering, hotels etc.

For our part, we will study the integration of the different proposals in the development of a system assistant in the field of computer device management. The main advantages of this new framework of methods [88] are that it requires less feature engineering process, less domain-specific wish list matching if the dataset is well- formed and from a single domain.

However, the model is still not very accurate. In this respect, it should be built with a structure in which fully natural language elements such as syntax, semantics, context, intention, etc., should be handled by the system in a more sophisticated way.

In the future, we plan to test the Seq2seq model in understanding more natural interactions. We also propose to reduce the complexity of the model for more success in real time and natural.

And from a development point of view, we are thinking of introducing reinforcement learning to increase the real-time capability of voice system assistants through experience.

Furthermore, it is interesting to mention that according to [91], the future field of application of the Chabot is the banking sector, finance, reality, real estate agents and homeowners.

Later, we plan to combine machine learning and NLP to design a new type of system assistant that can handle common hardware and software tasks based on natural interactions. In addition, sending text messages, taking calls, logging out of social networks, etc. will be possible.

Interface and voice assistant: New technologies

A look back at popular virtual assistants: Natural User Interfaces (NUIs) are supposed to be used by humans in a very logical way. However, the industry's attempt to deploy speech-based NUIs has had a significant impact on the naturalness of these interfaces.

A study [92] therefore carried out a functional and usability test of the most prestigious voice-activated personal assistants on the market, such as Amazon Alexa, Apple Siri, Microsoft Cortana and Google Assistant.

A comparison of the services each provides was also presented, taking into account access to music services, calendar, news, weather, to-do lists and maps/directions, among others. The test was designed by two experts in human-computer interaction and carried out by eight people.

The results show that, while there are many services available, there is still a lot of work to be done on the usability of these systems.

Furthermore, despite previous work to improve virtual assistants, speech recognition, contextual understanding and human interaction are still unresolved issues.

Thus, in order to focus and dissect these problems, 100 users participated in a research survey and shared their experiences [93].

According to the results, many services were covered by these assistants, but some improvements are still needed in speech recognition, contextual understanding and hands-free interaction.

A short time later, an article [94] describes the results of an evaluation of four intelligent personal assistants, to identify the best assistant based on the quality and correctness of their answers.

The results show that Alexa and Google are significantly better than Siri and Cortana. There is no statistically significant difference to confirm that Alexa is better than Google Assistant or vice versa.

Multimodal: With the recent launch of virtual assistant applications such as Siri, Google Now, S- Voice and Vlingo, oral access to information and services on mobile devices has become commonplace. These virtual assistants were, until now, limited mainly to voice input and output.

The Multimodal Virtual Assistant (MVA)

[95] is one of the first applications that allow users to plan an outing via an interactive multimodal dialogue with a mobile device. There, users can interact using combinations of voice and gesture inputs, and the interpretation of user commands depends on both the manipulation of the map and GUI 16 display

and the physical location of the device.

Building on this work, a paper [96] hypothesized that providing a higher level of visual and auditory immersion would improve the quality of the user experience.

Indeed, with the dramatic increase in computing power, the increased availability of 3D and immersive displays, and advances in artificial intelligence, the multimodal and intelligent virtual assistant is becoming much more practical. And it has been widely accepted that multimedia research should focus on quality of experience (QoE) as the primary measure for evaluating user experience.

Therefore, [96] focused on developing a scale to measure QoE and used it to evaluate the performance of the virtual assistant.

In order to test this hypothesis, the authors developed four variants of virtual assistant.

Because people have been communicating by voice for centuries, the introduction of the voice interface was initially thought to invoke the most natural human interaction with technology. However, designers of voice user interfaces are constantly constrained by the capabilities of their software and must design carefully due to the limited memory capacity of human users [97].

Usability is only one factor that influences our perception and intention to use a product experience. With only 70% of these devices being used at least once a week, the perceived usefulness of these voice experiences needs to be explored further.

To shed further light on this issue, a study [98] aimed to assess the extent to which cognitive load, relevance of visual information and personality influence the perceived usefulness of multimodal voice assistant technology in a within-subject repeated measures design.

In this study, as in previous studies, the results indicate that people perceive multimodal systems as more useful and perform better recall, when the voice assistant gives fewer answers and presents visual information on the screen that is relevant to the question the user has asked.

Interaction, architecture and control: As an intermediary, the virtual assistant will see all our personal data and control the providers we interact with. This raises many issues, including privacy, interoperability and generality.

As a result, a paper [99] presented the architecture of Almond, an open, participatory and programmable virtual assistant for online services and the Internet of Things (IoT) while preserving privacy.

This work addresses four challenges in virtual assistant technology including generality, interoperability, privacy and usability.

Users can ask Almond in natural language to perform any of the functions in the Thingpedia17 knowledge base. For privacy, all personal data and credentials are stored in the Thing System, whose code is open-source and can be run on personal phones or home servers.

Almond is the first virtual assistant able to convert rules with parameters expressed in natural language into code.

On the other hand, in order to overcome its privacy problems and all, a work has set the main objective of building a local voice assistant that does not depend on various technologies and cloud services, which would allow it to be used to solve various specific problems [100].

The study revealed that the creation and use of voice assistants is not only limited to cloud services but can also work in local development. In addition, the use of local systems expands the range of tasks in which they can be applied in IoT systems, smart home systems, healthcare, security and systems with a higher level of privacy, where the use of cloud technologies may be difficult.

Furthermore, in the imminent future, it cannot be ruled out that people are likely to engage with intelligent devices by teaching them natural language. A fundamental question to ask is how intelligent agents might interpret these instructions and learn new tasks.

In a pap [101], the authors present the first speech-based virtual assistant that can learn new commands through speech. In this work, they build on an earlier version of a text-based agent from [102], and have added support for a speech interface, as well as many different features, some of which are common to virtual assistants. An important observation is that the user study shows that people are enthusiastic about a personal agent that can learn new commands and the idea of sharing these taught commands with others.

Later, a significant contribution was made by a work [103] where the authors explored the influence of personalized voice assistant characters on user experience, acceptance, trust and workload compared to a non- personalized assistant.

Indeed, until recently, most assistants have lacked the level of interpersonal communication necessary to establish relationships. Related research suggests that in order to gain wider acceptance, these systems need to meet users' expectations of social interaction [104-105].

Therefore, they designed four assistant personas (friend, admirer, aunt and butler) and compared them to a (default) reference in a between-subjects study in real traffic conditions.

The results of this study show that personalization has a positive effect on trust and likeability if the voice assistant persona matches the user's personality.

Furthermore, with rapidly growing capabilities, increasing accuracy of voice recognition and an increasingly mature API over time, it is expected that the adoption of voice assistants will multiply in the next two years.

A thesis [106] explored the possibilities of integrating Kenotic Experience 18 with voice virtual assistants and implementing a voice integration module (KEVIN) for Kenotic Experience.

The implementation part was executed on three different areas: implementation on the Kenotic Experience side, implementation

of the Voice Hub application, implementation of voice applications for Alexa and the Google Assistant.

The main success of the implementation is that these three separate parts work together and form a single KEVIN functional module that is compatible with the two most popular voice assistant platforms.

By definition, Zero-UI [107] is a technology that uses our movements, our voice and even our thoughts to make the system react through our conditions. Instead of depending on clicking, dialing and tapping, customers will currently enter data by means of voice. Interactions will move away from phones and PCs to physical gadgets that we talk to.

Most current systems still have the disadvantage that only predefined voice directions are conceivable and that they can only store constrained commands.

With this in mind, a paper [108] is proposed with the aim of overcoming these drawbacks by making an autonomous personal assistant that can be associated exclusively by the client's voice. This research work aims to build a personal assistant using Raspberry Pi19 as the underlying processing chip and architecture.

Potential of recent intelligent assistants: The introduction of intelligent virtual assistants and the corresponding smart devices has brought a new degree of freedom to our daily lives. Voice-activated and internet-connected devices allow intuitive control and monitoring of devices anywhere in the world and define a new era of human- machine interaction.

Voice-based personal assistants have become so popular that we have opened up homes to devices like Amazon's Alexa, Google Home, Apple's Siri, Samsung's Viv, etc. Virtual voice assistants have great potential to disrupt the way people search for details and make this part of everyday conversation and changing lifestyles.

For further clarification, a paper [109] provided a brief introduction to voice support, installations, accuracy and adoption of voice technology by various service industries.

In this study, we can find information on the use of digital voice assistants that will be useful for academics as well as for business practitioners.

In the field of geo location, technology has long been helping tourists discover the city, from trip planning to find information on public transport, to navigational assistance and post-trip memories (photo sharing and online blogs).

Research [110] explored the utility of a hands-free, eyeless virtual tour guide, which could answer questions via a spoken dialogue user interface and inform the user of interesting features in view while guiding the tourist to various destinations.

The research highlighted the pleasure derived from this new form of interaction and revealed the complexity of prioritizing route guidance instructions alongside the identification, description and embellishment of landmark information.

Although virtual assistants are particularly successful in different domains, they also have great potential as artificial intelligence-based laboratory assistants.

Therefore, a work [111] aims at establishing a voice user interface for the control of laboratory instruments. The authors present an upgrade approach to integrate standard laboratory instruments with the Internet of Things (IoT).

This was achieved by using commercially available hardware, cloud services and open source solutions.

Later, a work [112] designed an application to help blind and visually impaired people to access library resources. Indeed, the development of smartphone technology from now on cannot be experienced by blind people.

It is created with the background application, where it will continue to run as long as the device is switched on.

In the medical sector, a lot of research into the intelligent virtual assistant has also been developed individually in recent years.

Sorting through this research and summarizing it, one study [113] concluded that a complete solution can only be achieved by combining these three points:

(i) Mobility - Movement and transport features from one place to another;

(ii) Physiological monitoring - to assess patients' physiological conditions and monitor their health status remotely;

(iii) Assisting with daily activities - assisting with self-care activities.

Subsequently, the authors decided to represent Docto-Bot [113], an autonomous humanoid biomedical robot, by maintaining these three parameters.

Moreover, with the propulsion of internet networks, a common concept nowadays is the replacement of face-to-face activities by the so-called "online", especially in the field of learning, where the Learning Management System (LMS) is considered as one of the most crucial elements of online activities.

Many organizations also use the LMS for their online activities (business, meetings, conferences, etc.) [114]. yet despite this high acceptance rate, many current studies [115–117] have highlighted usability and learning as a difficult research issue of current LMS.

As a result, a work [118] proposed an intelligent multi-agent search system for LMSs. The aim of this research was to investigate how a voice-activated virtual assistant affects the usability of an LMS during student learning activities. The results of the evaluation clearly indicate that the usability of the system is closely related to the ease of use of the system.

In addition to usability the ease of use of the system is therefore a very important factor for the user.

Due to the rise of Artificial Intelligence of Things (AIoT), Big Data 20, interactive technology and artificial intelligence, Smart Campus related applications are progressively valued from all horizons.

Moreover, the new mobile ecosystems (Android, iOS and Windows Phone) [119] allow us to use mobile services anytime and anywhere. This creates opportunities for new services, including those related to e-learning.

However, it is still rare to apply it in smart campus applications. There are not many examples presented in the form of catboats or personal mobile assistants.

One of these few works [120] presented NLAST, a system that functions as an assistant to students in their learning process. This assistant allows students to consult a repository of exam questions, to receive recommendations for learning material related to the exam questions they are examining, to ask questions about course content and to check their own assessed exams.

Following on from this work, a work [121] uses deep network technology to develop an emotionally sensitive Chabot and combined with the campus student affairs application to implement an emotionally sensitive, humanized and personalized campus virtual assistant.

In this study, the authors implemented an intelligent campus virtual assistant based on a deep convolution neural network (DCCN) and a long-short term memory recurrent neural network (RNN-LSTM).

Similarly, smart offices are dynamically changing spaces designed to improve employee efficiency, but also to create a healthy and proactive working environment.

Very recently, an article [122] presents the work undertaken to build the architecture necessary to integrate voice assistants into intelligent offices to support employees in their daily activities.

Very recently, an article presents the work undertaken to build the architecture necessary to integrate voice assistants into intelligent offices to support employees in their daily activities.

Recent challenges, research directorates and proposals on the voice assistant system

A virtual assistant can ultimately be our interface with all digital services. As an intermediary, the virtual assistant will see all our personal data and control the providers we interact with. It is therefore not surprising that all the big companies, from Amazon, Apple, Facebook, Google, to Microsoft, are competing to create the best virtual assistant.

In this section, we will analyses the challenges that today's virtual assistant systems still have to overcome, before exploring possible directions and presenting guidelines for future work.

Functionality and scope: To begin with, the demonstration in highlights incremental recognition, multimodal speech and gesture input, contextual language understanding and targeted clarification of potentially incorrect segments in the user's input.

Therefore, we propose to increase the performance of the virtual assistant command generation model by using multimodal learning, furthermore, by taking into account the time and space where the user is located.

In , we identified 3 major results:

Firstly, it is necessary to have a scale depending on a unique set of variables for each variant of virtual assistant to assess its QoE21.

Secondly, the immersive 3D visual output system emerged as the variant with the highest QoE. Users felt that the need for a virtual assistant in a virtual environment increases considerably.

Thirdly, the study shows a clear relationship between efficiency/accuracy and service quality assessment.

Furthermore, the main advantage of this scale is that the analysis tools can be used to assess even more dependent variables that will have an impact on quality of life.

Although there is still room for improvement, next steps include adding more functionality to the input/output system with more complex tasks, optimizing usability testing and testing with other visual feedback systems.

The best part is that this research is the beginning of new research and will play very important roles in the development of future virtual assistants. In particular, the knowledge provides promising data for the development of virtual assistants for virtual environments.

We therefore plan to use the QoE model to compare existing methods as well as to measure proposed new methods.

On the other hand, it is clear that systems that rely on a voice-only interface allow for more ease of user experience. According to the authors, the biggest problems were the quality of continuous ASR in an external environment, the management of large volumes of information, prioritization and the balance of push/pull information.

A first solution is therefore to design a speech recognition system capable of working well even in poor environments.

Second, to introduce an architecture into speech recognition and transcription that supports real-time speaker recognition [123], allowing NLP systems to accept only selected speech signals in near-real time to avoid any confusion in a multiparty conversation.

Furthermore, it is known that usability is only one factor that influences our perception and intention to use a product experience.

In, results showed that the attention demands of the task when the user performs tasks while interacting with the multimodal system can influence their ability to remember what the voice assistant has just said.

It is therefore recommended that UX 22 designers of the multimodal interface create succinct voice responses with audio-visual feedback for actions that need to be remembered.

And of course, the design of multimodal systems could improve the user experience and thus increase the use of voice interfaces.

In order to improve the synchronization between the virtual assistant and the user concerned, in addition to learned knowledge, introducing the notion of empathy is possible.

Development: Mentioned above, is the first virtual assistant able to convert rules with parameters expressed in natural language into code.

With a combination of a language and a menu-driven interface, Almond is ready to be used by enthusiasts to automate non-trivial tasks.

Two avenues for future directions have therefore emerged since this work. First, two design a new type of virtual assistant, which we will present the general concept in section 5, inspired by the Almond architecture. And secondly, to include the Thing Talk language in the system to encode the generated commands bringing more flexibility to the virtual assistant.

While there are many services available in popular virtual assistants, there is still a lot of work to be done to improve the usability of these systems.

Further research is needed. The potential of intelligent personal assistants could be tested in unexplored areas such as marketing, learning and sales. Furthermore, the combination of these devices and technologies with robots, data centers and machine learning techniques offers new opportunities.

With this in mind, several future directions came to mind, namely:

(1) Designing a more natural user interface but with much more precision.

(2) Find a new method of human- machine communication that is more natural, easy to adopt and user- friendly, even automatic.

(3) Broaden the scope of response by considering keywords as action clues for more response options.

(4) Use machine learning techniques in the field of speech interface.

(5) Facilitate the adoption of voice user interfaces by implementing them on mobile devices (local).

From a deployment point of view, emphasizes that the use of local systems extends the range of tasks where the use of cloud technologies may be difficult.

We therefore believe that leveraging local functionality to reduce cloud activity and remote data consumption would be a better fit for the design of a virtual assistant that can operate anywhere.

Literally, a good virtual assistant should be able to assist the actions of the user using it. Therefore, it must have the ability to control the user's activities well, both in the home automation domain and in other domains shows the general applicability of commercially available virtual assistants as laboratory assistants and could be of particular interest to researchers with physical disabilities or low vision.

The developed solution allows hands-free control of the instrument, which is a crucial advantage in the daily laboratory routine. The use of open communication protocols and data formats allows integration into available digital laboratory infrastructures. The use of this solution for the control of mobile devices is therefore crucial in the development of future virtual assistants.

In addition, a survey highlights users' experiences and uncovers some of the issues on which this discussion takes place:

(1) Speech recognition and contextual understanding problem

(2) Problem encountered in free human interaction.

Moreover, according to this survey, voice assistants have a rather low retention rate, with only 25% of frequent users in daily life. And yet, there is no shortage of use cases for virtual assistants.

There is a large proportion of people in the disabled community, mainly with cognitive impairments, who may have difficulty forming complete sentences and communicating, for whom the personal assistant can be a decisive factor.

Although there is room for improvement in all the devices tested, further studies can be conducted and the potential of intelligent voice assistants can be tested in various untouched areas such as education, banking, business, consulting, sales, etc.

And also a fusion of these devices with various machines, learning technologies and algorithms can give rise to various new possibilities.

Personally, instead of looking for a way to increase the number of frequent users, we should find a way to increase the internal use of the system (e.g. no explicit command, no restrictions, automated, free interaction, etc.). And then make it easier for the voice assistant to be used. (Good user experience)

Another important avenue is to exploit the power of deep learning both in understanding queries and in generating answers.

It is becoming essential to design informal language recognition systems for a more free and natural interaction. (stable to the non- respect of linguistic rules, multilingual).

Another work that deserves our attention for designing the virtual assistant of the future is [124]. Indeed, although SEVA is designed for system engineering, the idea is extensible to all domains where personal assistants are needed.

It is therefore possible to extend the idea to the field of virtual user interface assistants. A key point is also the design of several user assistants belonging to different user classes.

Furthermore, the results of an evaluation in

Clearly indicate that the usability of the virtual assistant system is closely related to the ease of use of the system. Indeed, students using LMS, serving as an experiment, with the proposed Scavenge approach showed a higher motivation and a high rate of task completion in a shorter time.

At the moment, the authors of this article are still working on extending the current version of Scavenge with more innovative initiatives and modern technological support.

Future research will focus on the development of a virtual bot for LMS that will have the ability to narrate and act on students' instructions, submit an assignment, update ongoing activities and perform several tasks on behalf of learners.

We propose the design of a new intelligent assistant system capable of acting on behalf of users. For this purpose, it is certainly more appropriate to include several agents in a multi-agent system.

Recently, the mobile intelligent assistant system can respond to all sorts of voice commands, send text messages, make phone calls, set reminders; anything we do on our phone, we can probably ask the virtual assistant to do for us.

Although according to [8], there are still limitations in these systems, for example: firstly, the virtual assistant does not speak some languages correctly. Secondly, when managing web applications, at some point the main thread is interrupted due to certain operations. Furthermore, the virtual assistant can only perform limited operations.

For this, we propose to increase the capacity of the virtual assistant through a new way of communication.

On the other hand, an advantage of using mobile devices is that we can, in addition, connect the virtual assistant with Android so that we can automate the computer and all devices connected with the virtual assistant with the Android application.

Therefore, we plan to install the virtual assistant on a mobile device so that we can connect with different devices.

In the future, machine learning and AI will be implemented in our design of the virtual assistant given the success of this field in the digital world.

In addition to the use of the mobile device for the installation, it is interesting to complement it with the IoT protocol in order to allow the virtual assistant a better administration of the digital objects in connection with the user.

Interaction, usability and user experience: Furthermore, the aim of a job is to facilitate access to information via a user- friendly interface. Chatter bots based on AIML23 are successful in these tasks. But they need to be redesigned for Android devices and adapted to specific functions.

Furthermore, the aim of a job is to facilitate access to information via a user- friendly interface. Chatter bots based on AIML are successful in these tasks. But they need to be redesigned for Android devices and adapted to specific functions.

The server platform also needs to be adapted to offer new functionalities.

In this respect, we plan to adapt an AIML for android to improve the usability of the relationship (naturalize) between man and machine.

Then in the future, it will be better to introduce the mixed approach (human- machine) in some requests to increase and adapt in real time the answer from the server.

In addition, listening to the user at all times also gives the user more confidence and comfort.

Creating the application with the background application, where it will continue to run as long as the device is switched on, would therefore make more sense.

Indeed, after analyzing the existing state of the art, the main satisfaction of users is ease of use and comfort before the different features.

Still a matter of experience, user research has also shown that people are enthusiastic about a personal agent that can learn new commands and share those taught commands with others.

In, the agent can be operated entirely remotely.

This work allows exploring a new horizon in virtual assistant systems. In the future, it is conceivable to add the ability to learn conditional execution to LIA.

However, in terms of deployment, it is still necessary to work on a lightweight version that only deals with responses that are not related to runtime commands.

In addition, the authors propose to explore a method for learning concepts (such as spam) through dialogue. Furthermore, once the user has given some explanation of the concept, if the agent cannot fully understand the whole explanation, it can identify some parts of the sentence that it does not understand and ask follow-up questions only on that part.

Although improving the performance of LIA without asking the user to explicitly mark whether an execution is correct or not is more than desirable.

To this end, we propose to exploit the mixed user/agent command to teach new commands to the machine. Second, we consider designing conditional execution for tricky commands returned by the virtual assistant. Logically, it is interesting to apply the approach to improve the execution of commands by the virtual assistant. Finally, in the design of the virtual assistant of the future, we should draw on concept learning to reinforce the commands already learned.

Another important point that should be studied is the personalization of voice assistants not only on a cultural level as it already happens, but on a context-sensitive basis. Thus, future voice assistants must adapt to the user as well as to the environment.

We are therefore obliged to adapt the behavior of the virtual assistant according to the contextual situation. For example, act quickly in an urgent situation and slowly but precisely in a calm situation.

It is interesting to create several types of personalities according to the type of user concerned.

Although several studies on virtual assistants have already been conducted [94], further studies are still needed to evaluate user interaction with intelligent personal assistants and to better understand how the interaction may affect the results obtained.

In contrast to, diverse populations can be included, which may strengthen the results.

Finally, an avenue worth exploring and which we will address in a thesis is the study of the feasibility of a new interaction

system: non- explicit interaction based on conversations between humans.

Furthermore, other aspects that could be the subject of future study are to explore different areas of use of the intelligent voice assistant in addition to proposing tools or frameworks to facilitate the operations and skills of new assistants. In addition, future work could examine the architecture models and implementation of the intelligent assistant following the expansion of the use of technologies such as wearable computing in addition to improving the user experience by considering usability, personalization and user behaviors.

Therefore, we plan to propose a new form of intelligent assistant (artificial co-user) which we will present the general concept in the last section. This necessarily implies a new architecture of a complete intelligent voice assistant system.

In order to gather more information on the use of the virtual assistant, very recently, a work carried out a usability assessment, based on a state-of-the-art literature review. This showed, firstly, that most users had never used a virtual assistant and even those who have such devices use them very rarely.

Secondly, the system is stable, especially for simple and moderate utterances, regardless of the user's experience. Finally, further analysis showed that the cause of the user's dissatisfaction is the system's inability to understand complex utterances.

In future work, the authors therefore propose to increase the control over the way virtual assistants work, taking into account that the user may pause for longer periods of time, or even wish to express enthusiasm.

In the context of natural language processing, they want to build models capable of predicting and classifying intentions and utterances based on unstructured input, mispronunciation, contradiction and exchanged words.

Similarly and in addition to these proposals, we plan to design a new comprehension system capable of predicting users' intentions (indirect commands) through other types of interactions, including conversations between colleagues, for example, to improve comprehension and provide more natural experiences in human-computer interaction [24].

Also, as the user rarely uses the intelligent voice assistant, we propose the introduction of a new more or less autonomous system allowing its use without direct human action.

Our main future work

Artificial co-user: New intelligent voice assistant exploiting human-to-human discussions

Context: Contacting devices with voice is now a common task for many. Voice assistants, for example Amazon Alexa, Microsoft Cortana, Google Assistant or Apple Siri, allow people, without having to own more mobile devices, to search for different topics, schedule a meeting or make a hands-free call from their car or home.

Several research papers have been published on the topic, including several surveys, since the introduction of the voice assistant in the 1990s. The type of voice assistant, the device and the user interface are presented in a general overview by explains the principles.

Assistance software, including ideas for a particular methodology and information- based address systems. The authors conclude that universal intelligent personal assistant software is relatively difficult to build.

We have previously carried out a partial review of the vast literature currently available on this topic in order to identify challenges, solutions and future perspectives on the field.

This led us to consider the feasibility of a new form of intelligent assistant, which we will briefly describe in this last section. Our main contribution focuses on the form of interaction between man and machine. In other words, the overall objective is to find a new means of interaction and communication to facilitate the distributed control and management of different machines or objects by an individual.

Obviously, we will be led to focus on the field of human-machine interface based on spoken languages which, visibly, is taking more and more its place in the attention of many researchers in the field of HMI, but also in natural language processing.

Indeed, as an innovative mode of interaction, the definition of the voice assistant is derived from advances in artificial intelligence, specialized systems, speech recognition, semantic websites, diagnostic systems and natural language processing.

1. **Issue:** Logically, modern interaction methods aim to provide its users with the best possible experience in the areas of communication and interaction. They are supposed to be highly accessible and easy to use. Despite the positive successes that natural language- based interaction methods have shown, they probably still have a gap in real-life practice, especially in natural conversation [125-126] as examples:

(1) In a conversation, the messages exchanged between individuals are necessarily floating and of poor quality that can be misinterpreted by the machine.

(2) Instructions can take many forms, direct or indirect, obvious or not, precise or not, etc.

(3) Humans can not only speak to each other using different languages, composing different languages at the same time, but also not respecting grammatical forms, which complicates processing.

(4) The machine takes a long time to respond to a request because of the delay in response on its part [127].

(5) When communicating between humans and several objects, hardware or software, confusing exchanges of instructions may occur [128].

In short, a question may arise: "How can we design an intelligent virtual assistant capable of assisting, or even directing, a user's instructions towards the computer objects he interacts with?"

In this perspective, we consider the design of a new form of virtual assistant that we will call "artificial co-user", co-user in the sense that the assistant will be able not only to perform the classical tasks of voice assistants but also should be able to take almost the place of the users. In other words, this co-user will act from time to time of its own accord but at the service of the user. Our motivation for this work also stems from the lesser success of virtual assistants in the real world.

Contributions: From the insights gained from related work, in addition to considerations of the perspectives and limitations of the research work in various recently published papers, we were able to come up with ideas that could advance research in the field of advanced human-computer interaction by highlighting the following research questions:

Question 1: By what means will the artificial co-user be able to predict instructions from complex sentences?

Question 2: How can confusion in the understanding of the co-user caused by the divergence of languages used by humans are avoided?

Question 3: By what means of communication can the user and the co-user exchange messages? And how can they always be connected in most cases?

Question 4: How can we achieve good synchronisation between the user and the co- user when generating an instruction?

Question 5: How can this user to co-user system work optimally?

These questions lead us to the following hypotheses, which we will try to validate throughout this work:

Hypothesis 1: A new approach to analysing keywords as a clue using machine learning allows for more generality when predicting appropriate orders.

Hypothesis 2: Introducing compound language analysis, in addition to French- English, should improve real-world practice in natural language processing.

Hypothesis 3: Mobile devices are the computing devices most likely to be present the majority of the time for men. In addition, they are already equipped with an audio sensor system [129].

Hypothesis 4: Recognition of the user's usual activities allows for more synchronization between the two controllers.

Hypothesis 5: This new human-computer interaction system can be applied in different types of environment. [130].

In carrying out this research, we proposed the following steps and solutions:

(1) First, we will introduce a compound language recognition model.

(2) Secondly, we will perform a feature selection and learning algorithm to extract key words in a conversation before generating the appropriate instructions from them.

(3)Third, we will install the agent in a mobile device that the individual uses the most.

(4) And finally, we will propose two interaction environments composed

(5) Of a human user, an artificial co-user, and several objects.

(6) We can see in the figure below an example of a simple illustration of the architecture of the two environments we propose. Indeed, by connecting the co-user to a mobile device, we should be able to automate the computer and all the devices connected with the virtual assistant **(Figure 2).**

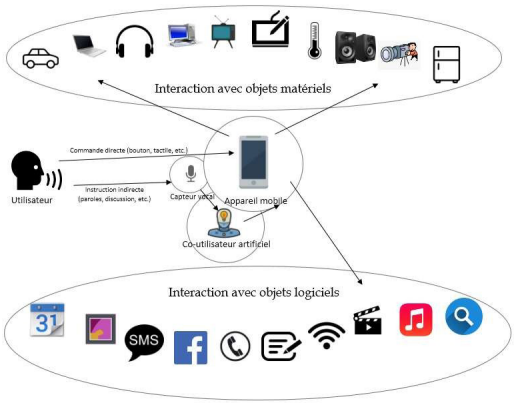

Figure 2 Architecture of the two interaction environments.

Conclusion

In recent years, new technologies and algorithms have enabled the development of many types of intelligent voice assistants. As a result, virtual assistants are becoming more and more common in our daily lives. Because of the good reach of smart phones, many of us have at least one on our mobile phones. These voice assistants offer a wide variety of features. However, this variety depends on the implementation and the purpose.

In this paper, we conduct a literature review to identify various critical areas of research on intelligent voice assistants in addition to discussing their concepts, architectures and components. To this end, we have conducted a systematic analysis of relevant articles from the last few years. We then discuss new technologies that have been developed in recent years. The analysis of the reviews has identified some gaps, trends, recent challenges, and open issues regarding the voice assistant system, including the components, before proposing research directions on the subject. The results of this study will provide information on the current and future state of the field of voice assistants that will be useful to academics as well as to new emerging researchers in the field.

In future work, we plan to study the feasibility of a new form of intelligent assistant that will focus on the form of interaction. This new concept should improve the success of virtual assistants in the real world, especially in poor conditions. Furthermore, in the context of natural language processing, we want to build models that can predict and classify intentions and utterances based on unstructured input, mispronunciation, contradiction and exchanged words. This concept is briefly described in Section 5.

In short, intelligent voice assistants have broader advantages for the future than they seemed to have. A fusion of devices with various machines and techniques should give rise to various new possibilities.

References

1 Azvine B, Djian D, Tsui KC, Wobcke W (2000) The intelligent assistant: An overview Intell Syst Soft Comput 20: 215-238.

2 POKOJSKI J (2004) Intelligent personal assistant in engineering activities. DS 34: Proceedings 7: 9-10.

3 Ricky MY, Gulo RS (2015) A Personal Agents in Ubiquitous Environment: A Survey. Procedia Comput Sci 59: 459-67.

4 Costa A, Julian V, Novais P, editors (2017) Personal Assistants: Emerg Comput Technol.

5 de Barcelos Silva A, Gomes MM, da Costa CA, da Rosa Righi R, Barbosa JL et al., (2020) Intelligent personal assistants: A systematic literature review. Expert Syst Appl 147: 113-193.

6 Rattan D, Bhatia R, Singh M (2013) Software clone detection: A systematic review Inf Softw Technol. 55: 1165-1199.

7 Roehrs A, Da Costa CA, da Rosa Righi R, De Oliveira KS (2017) Personal health records: a systematic literature review. J Med Internet Res 19: 56-59.

8 Chauhan K (2020) Virtual Assistant: A Review. Int j res eng sci manag 3: 138-140.

9 Pieraccini R, Director IC (2012) From AUDREY to Siri. Is speech recognition a solved problem? 23: 29-31.

10 Zambiasi SP, Rabelo RJ (2012) A proposal for reference architecture for personal assistant software based on soa. IEEE Lat Am Trans 10: 1227-1234.

11 Hauswald J, Laurenzano MA, Zhang Y, Yang H, Kang Y et al., (2016) Designing future warehouse-scale computers for sirius, an end-to-end voice and vision personal assistant. ACM Trans Comput Syst 34: 1-32.

12 Sarikaya R (2017) The technology behind personal digital assistants: An overview of the system architecture and key components. IEEE Signal Process Mag 34: 67-81.

13 Bellegarda JR (2013) Large-scale personal assistant technology deployment: the siri experience. In NTERSPEECH 13: 2029-2033.

14 Hsieh CH, Buehrer DJ (2014) The implementation of an artificially intelligent personal assistant for a personal computer. Appl Mech Trans Tech Public 627: 372-376.

15 Oishi S, Fukuta N (2016) A cooperative task execution mechanism for personal assistant agents using ability ontology. IEEE/WIC/ACM international conference on web intelligence (WI) 13: 664-667.

16 Pozna C, Foldesi P, Kovacs J (2013) The personal assistant application, problem definition. IEEE 4th International Conference on Cognitive Info communications 2: 851-856.

17 Yorke-Smith N, Saadati S, Myers KL, Morley DN (2012) The design of a proactive personal agent for task management. Int J Artif Intell Tools 21: 125-134.

18 Chihani B, Bertin E, Crespi N (2013) A user-centric context-aware mobile assistant. International Conference on Intelligence in Next Generation Networks (ICIN) 15: 110-117.

19 Cicció JA, Quesada L (2017) Framework for creating audio games for intelligent personal assistants. International Conference on Applied Human Factors and Ergonomics 17: 204-214.

20 Ponciano R, Pais S, Casal J (2015) Using accuracy analysis to find the best classifier for intelligent personal assistants. Procedia Comput Sci 52: 310-317.

21 Sun Y, Yuan NJ, Wang Y, Xie X, McDonald K et al., (2016) Contextual intent tracking for personal assistants. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 13: 273-282.

22 Bahrainian SA, Crestani F (2017) towards the next generation of personal assistants: systems that know when you forget. In Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval 1: 169-176.

23 Sun Y, Yuan NJ, Xie X, McDonald K, Zhang R (2017) Collaborative intent prediction with real-time contextual data. ACM Trans Inf Syst 35: 1-33.

24 Milhorat P, Schlögl S, Chollet G, Boudy J, Esposito A et al., (2014) Building the next generation of personal digital assistants. International conference on advanced technologies for signal and image processing 17:. 458-463.

25 Hauswald J, Laurenzano MA, Zhang Y, Li C, Rovinski A et al., (2015) Sirius: An open end-to-end voice and vision personal assistant and its implications for future warehouse scale computers. In Proceedings of the Twentieth International Conference on Architectural Support for Programming Languages and Operating Systems 14: 223-238.

26 Rybach D, Gollan C, Heigold G, Hoffmeister B, Lööf J et al., (2009) The RWTH Aachen University open source speech recognition system. InTenth Annual Conference of the International Speech Communication Association.

27 Seide F, Li G, Yu D (2011) Conversational speech transcription using context-dependent deep neural networks. InTwelfth annual conference of the international speech communication association.

28 Bay H, Tuytelaars T, Van Gool L (2006) Surf: Speeded up robust features. In European conference on computer vision 7: 404-417. Springer, Berlin, Heidelberg.

29 Jolad B, Khanai R (2019) An art of speech recognition: a review. In2019 2nd International Conference on Signal Processing and Communication (ICSPC) 29: 31-35.

30 Bahl LR, Brown PF, de Souza PV, Mercer RL (1989) A tree-based statistical language model for natural language speech recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 37:1001-1008.

31 Win KMN, Hnin Z Z, et YMKK (2020) Thaw, « REVIEW AND PERSPECTIVES OF NATURAL LANGUAGE PROCESSING FOR SPEECH RECOGNITION », Int J Res 1: 112-115.

32 Chugh A, Jerusha K, Krishnan etKS ,« A Review on Speech Recognition by Machines ».

33 Nassif AB, Shahin I, Attili I, Azzeh M, Shaalan K (2019) Speech recognition using deep neural networks: A systematic review. IEEE access. 7: 19143-191465.

34 Machine SR (2009) « A Review, MA Anusuya », Int J Inf, 6: 12-15.

35 Shahin I, Nassif AB, Hamsa S (2020) Novel cascaded Gaussian mixture model-deep neural network classifier for speaker identification in emotional talking environments. Neural Comput Appl 32: 2575-2587.

36 Nassif AB, Shahin I, Attili I, Azzeh M, Shaalan K (2019) Speech recognition using deep neural networks: A systematic review. IEEE access. 7: 19143-191465.

37 Wang P (2020) Research and Design of Smart Home Speech Recognition System Based on Deep Learning. International Conference on Computer Vision, Image and Deep Learning (CVIDL) 10: 218-221.

38 Panda SP (2017) Automated speech recognition system in advancement of human-computer interaction. International Conference on Computing Methodologies and Communication (ICCMC) 18: 302-306.

39 Xia L, Chen G, Xu X, Cui J, Gao Y (2020) Audiovisual speech recognition: A review and forecast. Int J Adv Robot Syst 17: 65-69.

40 Dupont S, Luettin J (2000) Audio-visual speech modeling for continuous speech recognition. IEEE transactions on multimedia. 2:141-151.

41 Singh J, Singh S, Vir D (2019) Classification of non-coding rna-a review from machine learning perspective. Life Sci. Inform. Publ.

42 Ganvir S, Lal N (2021) Automatic Speaker Recognition using Transfer Learning Approach of Deep Learning Models. In2021 6th International Conference on Inventive Computation Technologies (ICICT) Jan 20: 595-601.

43 Submitter I, Jena B, Mohanty A, et S. Mohanty K (2021) « Gender Recognition and Classification of Speech Signal », Bhagyalaxmi and Mohanty, Anita and Mohanty, Subrat Kumar, Gender Recognition and Classification of Speech Signal.

44 Kabir MM, Mridha MF, Shin J, Jahan I, Ohi AQ (2021) A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities. IEEE Access.

45 Soong FK, Rosenberg AE, Juang BH, Rabiner LR (1987) Report: A vector quantization approach to speaker recognition. Bell Syst tech j 66: 14-26.

46 Hyon S, Dang J, Feng H, Wang H, Honda K (2014) Detection of speaker individual information using a phoneme effect suppression method. Speech Communication. 57: 87-100.

47 Firoozeh N, Nazarenko A, Alizon F, Daille B (2020) Keyword extraction: Issues and methods. Nat Lang Eng 26: 259-291.

48 Marujo L, Habimana O, Li Y, Li R, Gu X et al., (2020) Sentiment analysis using deep learning approaches: an overview. Sci China Inf Sci 63: 1-36.

49 Joachims T (2002) Learning to classify text using support vector machines. Springer Science & Business Media.

50 Onan A, Korukoğlu S, Bulut H (2016) Ensemble of keyword extraction methods and classifiers in text classification. Expert Syst Appl 57:232-247.

51 Ni P, Li Y, Graham P, Nikolova N, Sankaran S (2020) Tension between leadership archetypes: Systematic review to inform construction research and practice. J Manage Eng 36: 311-319.

52 Biswas SK, Bordoloi M, Shreya J (2018) A graph based keyword extraction model using collective node weight. Expert Syst Appl .

53 Jones KS, Van Rijsbergen CJ (1976) Information retrieval test collections. J Doc.

54 Jain A, Mittal K, Vaisla KS (2020) FLAKE: Fuzzy Graph Centrality-based Automatic Keyword Extraction. J Comput.

55 Bordoloi M, Chatterjee PC, Biswas SK, Purkayastha B (2020) Keyword extraction using supervised cumulative Text Rank. Multimed Tools Appl 79: 31467-31496.

56 Hasan KS, Ng V (2014) Automatic keyphrase extraction: A survey of the state of the art. In Proceedings of the 52nd Annual Meeting of the Assoc Comput Linguist 1: 1262-1273.

57 Papagiannopoulou E, Tsoumakas G (2020) A review of keyphrase extraction. Wiley Interdisciplinary Reviews: Data Min Knowl Discov .

58 Sun C, Hu L, Li S, Li T, Li H et al (1864) A Review of Unsupervised Keyphrase Extraction Methods Using Within-Collection Resources. Symmetry. 12: 18-64.

59 Hasan KS, Ng V (2010) Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art. In Coling 10: 365-373.

60 Fellbaum C (1998) « A semantic network of English verbs », WordNet: An electronic lexical database 3: 153-178.

61 Wan X, Yang J, Xiao J (2007) Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. InProceedings of the 45th annual meeting of the Assoc Comput Linguist 7: 552-559.

62 Romo JM, Araujo L, Fernandez AD (2016) S em G raph: Extracting keyphrases following a novel semantic graph-based approach J Assoc Inf Sci Technol 67: 71-82.

63 Alrehamy H, Walker C (2018) Exploiting extensible background knowledge for clustering-based automatic keyphrase extraction. Soft Comput 22: 7041-7057.

64 Liu Z, Li P, Zheng Y, Sun M (2009) Clustering to find exemplar terms for keyphrase extraction. InProceedings of the 2009 conference on empirical methods in natural language processing 9: 257-266.

65 Rabby G, Azad S, Mahmud M, Zamli KZ, Rahman MM (2020) Teket: a tree-based unsupervised keyphrase extraction technique. Cognitive Computation. 12: 811-833.

66 Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding.

67 Peters ME, Neumann M, Iyyer M, Gardner M, Clark C et al., (2018) Deep contextualized word representations.

68 Lee J, Yoon W, Kim S, Kim D, Kim S et al., (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36: 1234-1240.

69 Beltagy I, Lo K, Cohan A (2019) Scibert: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676.

70 Sahrawat D, Mahata D, Zhang H, Kulkarni M, Sharma A et al., (2020) Keyphrase extraction as sequence labeling using contextualized embeddings. Adv Inf Retr 12:328-330.

71 Huang Z, Xu W, Yu K (2015) Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.

72 Zhang H, Long D, Xu G, Xie P, Huang F et al., (2020) Keyphrase Extraction with Dynamic Graph Convolutional Networks and Diversified Inference. arXiv preprint arXiv:2010.12828.

73 Han D, Song X, Cui Y (2020) An Extractive Chat Summary Generation Method for Ecommerce Chatbots. DEStech Trans Comput Sci Eng.

74 Behere T, Vaidya A, Birhade A, Shinde K, Deshpande P et al., (2020) « Text Summarization and Classification of Conversation Data between Service Chatbot and Customer », in 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4, 20: 833-838.

75 Scherbakova A (2020) Comparative Study Of Data Clustering Algorithms And Analysis Of The Keywords Extraction Efficiency: Learner Corpus Case. Higher School of Economics Research 20: 86-97.

76 Pikies M, Riyono A, Ali J (2020) Novel Keyword Extraction and Language Detection Approaches. arXiv preprint arXiv:2009.11832.

77 Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR et al., (2019) Generalized autoregressive pretraining for language understanding. Adv Neural Inf Process Syst 32: 45-49.

78 Cer D, Yang Y, Kong SY, Hua N, Limtiaco N et al., (2018) Universal sentence encoder. arXiv preprint arXiv:1803.11175.

79 Zhu X, Lyu C, Ji D, Liao H, Li F (2020) Deep neural model with self-training for scientific keyphrase extraction. Plos one 15: 34-39.

80 Gagliardi I, Artese MT (2020) Semantic unsupervised automatic keyphrases extraction by integrating word embedding with clustering methods. Multimodal Technol Interact 4: 28-30.

81 Wang J, Su G, Wan C, Huang X, Sun L (2020) A Keyword-Based Literature Review Data Generating Algorithm—Analyzing a Field from Scientific Publications. Symmetry 12: 901-903.

82 Grassi L, Recchiuto CT, et A Sgorbissa (2020) « A Knowledge-Based Conversation System for Robots and Smart Assistants », PhD Thesis, University of Genoa.

83 Zhong P, Liu Y, Wang H, Miao C (2021) Keyword-Guided Neural Conversational Model. InProceedings of the AAAI Conference on Artificial Intelligence 35: 14568-14576.

84 Zhang S, Dinan E, Urbanek J, Szlam A, Kiela D et al., (2018) Personalizing dialogue agents: I have a dog, do you have pets too?

85 Shukla R (2004) Keywords Extraction and Sentiment Analysis using Automatic Speech Recognition 7: 34-38.

86 Ekbal A (2020) Towards building an affect-aware dialogue agent with deep neural networks. CSI Transactions on ICT 8: 249-255.

87 Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. Adv Neural Inf Process Syst 14: 3104-3112.

88 Prassanna J, Nawas KK, Jackson C , Prabakaran R, et S. Ramanat, « Towards Building A Neural Conversation Chatbot Through Seq2Seq Model .

89 Zhang W (2020) Intelligent Personal Assistant Dialog Generation using Paraphrasing.

90 sampling, Fu Y, Feng Y, Cunningham JP (2020) Paraphrase generation with latent bag of words. arXiv preprint arXiv:2001.01941.

91 Pugalenthi R, Chakkaravarthy AP, Ramya J, Babu S, et R. R. Krishnan (2020) « Artificial learning companionusing machine learning and natural language processing », Int J Speech Technol 20: 1-8.

92 López G, Quesada L, Guerrero LA, Alexa vs. Siri vs (2017) Google Assistant: a comparison of speech-based natural user interfaces. In International Conference on Applied Human Factors and Ergonomics 17: 241-250.

93 Tulshan AS, Dhage SN (2018) Survey on virtual assistant: Google assistant, siri, cortana, alexa. Int Symp Intell Signal Process Commun Syst 19: 190-201.

94 Berdasco A, López G, Diaz I, Quesada L, Guerrero LA (2019) User experience comparison of intelligent personal assistants: Alexa, Google Assistant, Siri and Cortana. In Multidisciplinary Digital Publishing Institute Proceedings 31: 1-51.

95 Johnston M, Chen J, Ehlen P, Jung H, Lieske J et al., (2014) Mva: The multimodal virtual assistant. In Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue 14: 257-259.

96 Saad U, Afzal U, Issawi AE, Eid M (2017) A model to measure QoE for virtual personal assistant. Multimed Tools Appl 76: 12517-12537.

97 Bigot L, Caroux L, Ros C, Lacroix A, Botherel V (2013) Investigating memory constraints on recall of options in interactive voice response system messages. Behav Inf Technol 32: 106-116.

98 Baeza RR, Kumar AR (2019) Perceived Usefulness of Multimodal Voice Assistant Technology. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting 63: 1560-1564.

99 Campagna G, Ramesh R, Xu S, Fischer M, Lam MS (2017) Almond: The architecture of an open, crowdsourced, privacy-preserving, programmable virtual assistant. InProceedings of the 26th International Conference on World Wide Web 3: 341-350.

100 Polyakov EV, Mazhanov MS, Rolich AY, Voskov LS, Kachalova MV et al., (2018) Investigation and development of the intelligent voice assistant for the Internet of Things using machine learning. Electron Netw Technol 14: 1-5.

101 Chkroun M, Azaria A (2019) Lia: A virtual assistant that can be taught new commands by speech. International J Hum Comput Int 35: 1596-1607.

102 Azaria A, Krishnamurthy J, Mitchell TM (2016) Instruct able intelligent personal agent. In Thirtieth AAAI conference on artificial intelligence.

103 Braun M, Mainz A, Chadowitz R, Pfleging B, Alt F (2019) At your service: Designing voice assistant personalities to improve automotive user interfaces. InProceedings of the 2019 CHI Conference on Human Factors in Computing Systems 2: 1-11.

104 Nass C, Jonsson IM, Harris H, Reaves B, Endo J et al., (2005) Improving automotive safety by pairing driver emotion and car voice emotion. InCHI'05 extended abstracts on Human factors in computing systems 13: 1973-1976).

105 Malle BF, Magar ST (2017) What kind of mind do I want in my robot? Developing a measure of desired mental capacities in social robots. InProceedings of the companion of the ACM/IEEE international conference on human-robot interaction 6: 195-196.

106 Kuzmin BD (2021) Kentico Voice Interface.

107 Mane P, Sonone S, Gaikwad N, Ramteke J (2017) Smart personal assistant using machine learning. International Conference on Energy, Communication, Data Analytics and Soft Computing 17: 368-371.

108 Vashistha P, Singh JP, Jain P, Kumar J (2019) Raspberry Pi based voice-operated personal assistant (Neobot). International conference on Electronics, Communication and Aerospace Technology (ICECA) 12: 974-978.

109 Darda P, Chitnis M, « A Review on Voice Assistant Adoption in Service Sector ».

110 Bartie P, Mackaness W, Lemon O, Dalmas T, Janarthanam S et al., (2018) A dialogue based mobile virtual assistant for tourists: The Space Book Project. Computers, Comput Environ Urban Syst 67: 110-123.

111 Austerjost J, Porr M, Riedel N, Geier D, Becker T et al., (2018) Introducing a virtual assistant to the lab: A voice user interface for the intuitive control of laboratory instruments. slas technology: Translating Life Sciences Innovation 23: 476-482.

112 Barata M, Salman AG, Faahakhododo I, Kanigoro B (2018) Android based voice assistant for blind people. Library Hi Tech News 6: 13-19.

113 Hossain MA, Qureshi MJU (2021) IoT Based Medical Assistant Robot Int. J Electr Comput Eng.

114 Laeeq K, Memon ZA (2018) An Integrated Model to Enhance Virtual Learning Environments with Current Social Networking Perspective. Int J Emerg Technol Learn 13: 15-18.

115 Bartolotta J, Newmark J, Bourelle T (2018) Engaging with online design: Undergraduate user-participants and the practice-level struggles of usability learning. 65: 63-72.

116 Harris HS, Greer M (2017) Over, under, or through: Design strategies to supplement the LMS and enhance interaction in online writing courses. Communication Design Quarterly Review. 4: 46-54.

117 Oliveira PC, Cunha CJ, Nakayama MK (2016) Learning Management Systems (LMS) and e-learning management: an integrative review and research agenda. J Inf Technol Manag 13: 157-180.

118 Laeeq K, Memon ZA (2021) Scavenge: An intelligent multi-agent based voice-enabled virtual assistant for LMS. Interact Learn Environ 29: 954-972.

119 Joseph J, Shinto Kurian K (2013) Mobile OS–Comparative Study. J Eng Appl Sci 2: 10-19.

120 Fonte FA, Nistal ML, Nistal ML, Rodríguez MC (2016) NLAST: A natural language assistant for students. In2016 IEEE global engineering education conference (EDUCON) 10: 709-713.

121 Chiu PS, Chang JW, Lee MC, Chen CH, Lee DS (2020) Enabling intelligent environment by the design of emotionally aware virtual assistant: A case of smart campus. IEEE Access. Int j innov technol expl eng 8: 36-41.

122 Bogdan R, Tatu A, Crisan-Vida MM, Popa M, Stoicu-Tivadar L (2021) A Practical Experience on the Amazon Alexa Integration in Smart Offices. Sensors 21: 730-734.

123 Damacharla P, Dhakal P, Bandreddi JP, Javaid AY, Gallimore JJ (2020) Novel Human-in-the-Loop (HIL) Simulation Method to Study Synthetic Agents and Standardize Human–Machine Teams (HMT). Applied Sciences 10: 83-90.

124 Krishnan J, Coronado P, Reed T (2019) SEVA: A Systems Engineer's Virtual Assistant. In AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering.

125 Oukrich N (2019) Daily Human Activity Recognition in Smart Home based on Feature Selection, Neural Network and Load Signature of Appliances.

126 Lee C, Han D, Jin H, Oh A (2019) automaTA: Human-Machine Interaction for Answering Context-Specific Questions. InProceedings of the Sixth (2019) ACM Conference on Learning@ Scale 24: 1-4.

127 Rea F, Vignolo A, Sciutti A, Noceti N (2019) Human motion understanding for selecting action timing in collaborative human-robot interaction. Front Robot AI6: 52-58.

128 Schaub LP, Vaudapiviz C (2019) Les systèmes de dialogue orientés-but: état de l'art et perspectives d'amélioration Goal-oriented dialog systems: a recent overview and research prospects. In Actes de la Conférence sur le Traitement Automatique des Langues Naturelles 3: 541-562.

129 Bâce M, Staal S, Bulling A (2020) How far are we from quantifying visual attention in mobile HCI?. IEEE Pervasive Comput 19: 46-55.

130 Chao YW (2019) Visual Recognition and Synthesis of Human-Object Interactions (Doctoral dissertation) Univ Michigan Lib.