A VIRTUAL ASSISTANT FOR WEBSITES

Lizandro Kirst da Silva, Daniel Ribeiro Brahm, Gustavo Tagliassuchi
ADS Digital
Email: lizandro@adsdigital.com.br, daniel@adsdigital.com.br, gustavo@adsdigital.com.br

José Luiz Duizith
Lutheran University of Brazil – ULBRA, Brasil
ADS Digital
Email: jose@adsdigital.com.br

Stanley Loh
Lutheran University of Brazil – ULBRA, Brasil
Catholic University of Pelotas – UCPEL, Brasil
E-mail: sloh@terra.com.br

Feliz Gouveia
University Fernando Pessoa / CEREM, Portugal
Email: fribeiro@ufp.pt

Keywords: virtual assistant, intelligent interfaces, chatterbots, natural language interaction

Abstract: This work presents a Virtual Assistant (VA) whose main goal is to supply information for Websites users. A VA is a software system that interacts with persons through a Web browser, receiving textual questions and answering automatically without human intervention. The VA supplies information by looking for similar questions in a knowledge base and giving the corresponding answer. Artificial Intelligence techniques are employed in this matching process, to compare the user's question against questions stored in the base. The main advantage of using the VA is to minimize information overload when users get lost in Websites. The VA can guide the user across the web pages or directly supply information. This is especially important for customers visiting an enterprise site, looking for products, services or prices or needing information about some topic. The VA can also help in Knowledge Management processes inside enterprises, offering an easy way for people storing and retrieving knowledge. An extra advantage is to reduce the structure of Call Centers, since the VA can be given to customers in a CD-ROM. Furthermore, the VA provides Webmasters with statistics about the usage of the VA (themes more asked, number of visitants, time of conversation).

1. INTRODUCTION

With the growing popularity of the Web (World Wide Web), a great number of websites is available for people visit. In the same way, information available inside these sites is increasing, generating the information overload. This problem occurs when people have too much information so that they can not find what is desired, leading to lost users (CHEN, 1994).

To minimize these problems, websites and portals offer a list of FAQs (*frequently asked questions)*, containing pairs question-answer that correspond to information usually searched by users. However, FAQs also have problems. Even in a medium size list, it is difficult for the user to find the question more similar to his/her need.

A similar problem also happens inside organizations. People need to share knowledge, but there are few resources to support this kind of collaborative process. The consequence is lost knowledge (when people leave the organization) or re-work (two or more persons searching for the same information).

Furthermore, problems will persist if the organization does not have a way to store knowledge, so that people can reuse or disseminate it.

Knowledge Portals or Corporate Portals intend to reduce these problems. They are websites accessible internally by organization staff, where explicit knowledge is available for storage and retrieval. However, knowledge has to be structured in databases and this is not a natural task for people not familiar with technologies.

In addition, even portals may grow unstructured or may cause information overload due to the huge volume of information available for the users.

This work presents a Virtual Assistant (VA), a software that receives a question in free natural language and answers querying a knowledge base. In this base, knowledge is formatted as question-answer pairs. This facilitates its storage by people. When a user puts a question to the Assistant, it looks in the base for the more similar question and returns the respective answer. This facility allows the users to enter questions in natural language, without restrictions and without having to seek in lists or to use query languages as in databases.

The interaction between the VA and the user happens naturally and intuitively as in a Web chat.

This paper is structured as follows. Section 2 discusses knowledge management actions and how VAs can help these processes. Section 3 discusses differences between VAs and chatter bots. In the section 4, the VA is presented and its features and functions detailed. Section 5 explains how the VA was developed. Section 6 presents some cases where the VA is being used and discusses a special situation where the VA is used for marketing. Section 7 presents concluding remarks and discusses future directions.

2. KNOWLEDGE MANAGEMENT AND VIRTUAL ASSISTANTS

Learning organizations are those that use knowledge as one of the main resources (SENGE, 2001). They know how to use knowledge for improving business, and they know that knowledge, as other resources, need to be acquired and managed. Knowledge Management (KM) is the area responsible for capturing, storing and retrieving knowledge to support business processes (DAVENPORT, & PRUZAC, 1997).

According to NONAKA & TAKEUCHI (1995), the majority of the organizational knowledge comes from interactions between people. People tend to reuse solutions from other persons in order to gain productivity.

Collaboration is one of the most important tasks for innovation and competitive advantage within *learning organizations* (SENGE, 2001). It is important to record knowledge to later reuse. If knowledge is not adequately recorded, organized and retrieved, the consequence is re-work, low productivity and lost of opportunities.

Virtual Assistants (VAs) can help in knowledge management processes. Using a standard format (questions and answers), people can easily store knowledge. This can encourage organizational staff to explicit their tacit knowledge. Thus, organizational knowledge tends to increase in explicit formats.

Other advantage is that VAs facilitate the retrieval of knowledge, allowing people to search information through questions in natural language and freeing them from searching web pages, FAQ lists or to learn artificial languages or signals.

Furthermore, the knowledge base may gradually increase, because each question that was not answered is stored for future analysis.

3. VIRTUAL ASSISTANTS AND CHATTER BOTS

Software that interacts with users in natural language is not new. The most famous one is Eliza (WEIZENBAUM, 1967), that simulates a psychiatric session.

In the Web, there are many of this kind and they are called chatter bots (or chatter robots), because they simulate a Web chat with users.

The majority of the chatter bots only have as goal to maintain the conversation coherently. Answers may be evasive and not necessarily the user's question is answered. The success is to pretend to be a human interlocutor (as in the Turing test).

Although the embedded technology, these systems are not suited to supply information. They serve as marketing resource, helping in the branding process (to spread or to strengthen the trademark).

On the other side, there are Virtual Assistants (VA). They have to supply users with information. The responsibility is to answer the user's question with the correct information.

Its goal is not to simulate a human. Even because human conversation has interjections, incomplete phrases, connotations, subject variations and so on, which can obstruct the conversation goal (ZUE, 1997).

In this sense, a VA does not have to answer all kinds of question (for example, general culture). Its knowledge base is limited to a domain.

An example of a VA is discussed in ZUE (1997). The Jupiter system allows conversation by telephone about weather forecasting in some cities. The system has a vocabulary of 1500 words. First the system recognizes spoken words, after it mounts a context and then generates a paraphrase in SQL (standard language for relational databases). Finally, the system queries a database and generates a locution (voice synthesis) corresponding to the answer.

4. THE PROPOSED VIRTUAL ASSISTANT (VA)

The Virtual Assistant (VA) presented in this paper receives a question written in unrestricted natural language posted by a user using a Web browser. The VA analyzes the question, eliminating stopwords (generic words like prepositions and articles) and reducing words to their radicals, using a stemming algorithm.

After that, the VA looks in the knowledge base for similar questions and returns to the user the corresponding answer. Knowledge is stored as a set of question-answer pairs. The knowledge base is maintained by domain experts, responsible by storing knowledge as questions-answers and by updating the base (increasing the stored knowledge).

The facility is that the question does not have to be exact. For example, storing "*what is a virtual assistant*", the VA can answer questions like "*please tell me what is a virtual assistant*" and "*can you answer me what a virtual assistant is ?*".

The matching algorithm uses a similarity function, based on Artificial Intelligence and Natural Language Processing methods, that analyzes stored questions applying combinatorial methods (besides the stemming algorithm). Thus, storing "*product reduce cost*", the VA can answer questions like "*does the product reduce costs*" or "*is it true that we can reduce costs using the product ?*" or "*can we get cost reduction using the product ?*".

Furthermore, the VA uses minimal combinations to answer questions. So, it is not necessary to store complete or exact questions. For example, the stored question "*cost product*" is similar to "*what is the cost of the product*" and to "*how much does the product cost*".

The method is based on the assumption that users tend to use direct and short questions or phrases. A study of (ZUE, 1997) concluded that customers of an organization utilized short phrases (12 words or less), during interactions via telephone. Our premise is that, in a written channel, users will use less words.

In the same study, ZUE (1997) noted that 70% of the phrases contained only words of a predefined vocabulary, 11% of the phrases had words not present in the vocabulary, 11% of the utterances do not have any word (silence or laughs) and 8% were not recognized phrases. This confirms our premise that a sub-set of the natural language (called domain vocabulary) is supposed to answer the majority of the questions in an application. Thus, experts can store a limited set of knowledge for answering a great variety of users´ questions.

Answers may be stored in HTML format. This allows including links to web pages or to send e-mails messages ("mailto").

Furthermore, it is possible to associate to each answer an image (picture, graphic, photo, cartoon, that is, a GIF or JPG file). This allows presenting a product or how to use, or even to associate a human figure in order to pass emotion or to act as a human is interacting with the user (smiling, pointing, showing).

An additional service is that synonyms can also be stored in the knowledge base. The VA has a module that allows the user to manage synonyms (to set a word as the standard and to associate a list of synonyms).

The VA also recognizes differences between "*what is an assistant*" and "*what is the price of an assistant*", using an intelligent method for analyzing significant words. The VA identifies one significant word in the first question ("*assistant*") and two significant terms in the second ("*price*" and "*assistant*"). Thus, it is possible to store two different questions "*what is an assistant*" and "*price assistant*" so that no confusion will be made when answering the questions in the begin of this paragraph.

The VA allows storing two or more answers for the same question. In this case, the same answer must be stored twice, each time with a different answer. When this question is asked by the user, the VA selects randomly one of the answers. This allows the VA not to be monotonous.

At the moment, the VA has support only for the Portuguese language. A engine is being developed to hold questions written in English.

4.1 Other functions of the VA

The VA has an administrative module, where other functions are available.

The administrative module has functions to list what was stored (questions and answers) and to edit or remove knowledge.

There is a module to store standard-answers. These answers will be presented to the user when the VA does not find a similar question. This is useful to give the user a plausible answer, instead of giving a wrong answer or showing nothing.

For example, assuming there are two questions stored as "*product price*" and "*product working*", if the user asks "tell me about the product", the VA will answer with "*please, be more specific about what you want about product*".

A special function of the VA is to store non-answered questions as unknown questions. That is, when the VA does not find a similar question in the knowledge base, it stores the user´s question in a list of unknown questions. These questions can be analyzed later by experts in order to insert new questions in the base. This way, the knowledge base can increase along the time and the experts do not have to store all the possible questions once at a time, before the initial use of the VA.

Other service provided by the VA is the offer show. Offers are utterances of the VA when the user does not ask anything. After a predefined time period of "silence", the VA presents one of stored offers. Offers may be used for publicity, showing new products or promotions, or it may be used to guide the user in how to ask questions.

A special function allows listing all sessions and the conversations performed between the user and the VA. This is useful to understand how the VA is answering questions.

A statistical module allows analyzing which answers are the most presented and which words are the most used by users inside questions. This is important to understand what users are looking for (themes or subjects) and what kind of language they are using.

This statistical module also gives information about the sessions. For example, it is possible to obtain the minimum, medium and maximum time of conversation. This is important to understand how much time users are spending when talking with the VA. It is also possible to obtain the number of accesses (different sessions) during a given time period (in days). This is useful to analyze the popularity of the VA and in which days the VA was most used.

5. IMPLEMENTATION

The Virtual Assistant (VA) was implemented using free software technologies (*open source*), like PHP, Javascript and MySQL.

Javascript is a programming language for Web-based systems that provides better interactivity with users. Most part of the application is manipulated in the web browser without the need of a server, as for example form and data validation, presentation of new windows and image manipulation. The majority of the web browsers support Javascript.

PHP is a programming language used in web servers. It allows dynamic pages and improves interaction with a set of complex operators (CHOI et al., 2001).

MySQL is an open source database management system that supports SQL standard. It enables reliable and efficient applications, besides its facilities to create, manipulate and administrate data.

5.1 Techniques used in the VA

The main goal of the VA is to hold sentences written in unrestricted natural language. This means users can use incomplete phrases or even non-well formed sentences.

The first step is to remove the *stopwords*. According to SALTON & McGILL (1983), *stopwords* are too much frequent terms, used in almost all texts and having a non-specific meaning (e.g., prepositions and articles). *Stopwords* are removed because they do not influence the analysis of the meaning.

After that, the VA treats synonyms, according to a list previously created by an administrative user (see section 4).

The VA also uses a stemming algorithm for reducing words to radicals. This algorithm analyzes the origin of the words in order to understand and hold morphological variations.

The assumption is that a subject or concept may be identified by a group of correlated words with the same origin. For example, "*product*", "*production*" and "*produce /produces / produced / producing*" lead to the same idea.

The stemming technique allows reducing the size of an index and accommodating variations used by different users to express the same idea.

The VA uses a modified version of the Paice/Husk algorithm adapted to portugese by ORENGO & HUYCK (2001).

To find the most similar question, the VA compares the question input by the user against all questions stored in the knowledge base. The similarity function matches questions using words present in both questions.

6. CASES AND APPLICATIONS

Virtual Assistants may be used in different applications, being enough to change the knowledge base.

The Brazilian university ULBRA (www.ulbra.br) utilizes a VA to guide internal collaborators in the Intranet/Corporate Portal. The VA answers questions like where to find something or how to fulfill a certain electronic form. Besides that, the VA is used to orient people in the total quality program, supplying information about rules and best-practices. The main advantage of using a VA in this context is to reduce internal calls to some departments and to disseminate organizational knowledge, without collaborators having to attend to courses or to take time from colleagues. Furthermore, the statistical results extracted by the VA help to understand what kind of doubts and problems are more frequent. This kind of analysis helps the university staff to plan new training programs.

The Plug-In Data Center (www.plugin.com.br) has a VA in its internal department of Technical Support. The VA is used by authorized customers and its function is to explain the working of the data center and how technical operations should be performed. Besides that, the VA supports human attendees when answering questions from online customers by phone. The VA has information about prices, conditions and promotions. This part of the VA will be open to non-customers soon, in order to help the sales department in acquiring new customers. The main advantage of the VA is to reduce the infra-structure needed to attend customers and the time for training human attendees, especially technical ones. Statistical analysis of the most frequent words used in the questions have helped to find which promotions have received more questions about. This helped to maintain successful strategies.

Nucletec is an enterprise that produces health equipments. At the moment, it uses a VA to answer questions about the PipiStop, a device that helps in the treatment of children with nocturnal enuresis (bedwetting). The VA answers questions as where and how to buy one, how to use it in children, how to change the battery and so on (www.pipistop.com.br).

Other kinds of applications of the VA include:

Websites of Technical Support: a VA can help to reduce costs with support services; customers will be oriented to first ask their questions to the VA in a Website; besides that, the VA can be put into a Compact Disk, so that customers will have support 24x7 in own home;

Electronic commerce sites: a VA can be used to give potential customers information about products and services, besides explaining advantages and how to use them;

E-learning and remote training: a VA can be enriched with content about subjects in order to “teach” students;

CRM systems: VAs can interact with Customer Relationship Management systems in remote touch points, instead of using human attendees.

6.1 Virtual Assistants for Marketing

One case where the VA is being used for marketing is the web portal of the Third Tuesday (www.terceiraterca.com.br), an event to enable business networking. The VA answers questions about the event (what it is, where and when it happens, who participates in).

But the main benefit of the VA is for marketing. In this portal, the VA acts as a chatterbot, answering also generic questions.

The interesting in this VA is that the front character of the VA is a virtual woman, called Tetê, created to be the hostess of the event. She has her own personality and culture and does not tolerate cheek. This VA entertains visitants of the portal and helps to divulgate the event by viral marketing (one person saying to another) and by strengthening the mark. This VA receives 20 visitants a day and each person spends a medium of 4 minutes talking with the VA. Some people have spent almost 1 hour interacting with the VA. The knowledge base is composed of more than 1,500 questions.

Special functions were implemented in this VA to give daily information about currency exchange and weather forecasting. This information is captured automatically and online from other websites.

The VA Tetê also helps in divulgating other enterprises (partners) through the offering mechanism.

A similar VA is being created to divulgate tourism attractions in the brazilian southern state of Rio Grande do Sul (www.adsdigital.com.br/cases). This VA has photos about cities, events and places to visit. Yet, a little of the history of this state is explained by the VA through historical places and museums.

The difference in this case is that the VA guides the user, asking questions and so offering the attractions according to the answers of the user. This allows understanding preferences of a person and then selecting suited items.

7. CONCLUSIONS

This work presented the technology of Virtual Assistants (VAs) for answering natural language questions. The benefits of using VAs are broad, including to reduce costs with customer support and service and to do marketing through entertainment.

However, the main advantage of a VA is to help in the Knowledge Management process. With the VA, people can easily store and retrieve knowledge, through question-answer format. For example, in the Third Tuesday case (Tetê), 300 question-answer pairs were stored in about 3 hours (along 2 days).

Although the benefits, some limits exist. One is that the quality of the answers depends on the quality of the knowledge base. Few questions in the base will produce too much standard answers (those presented when the VA does not find a similar question in the base). This may discourage users; they may get frustrated and do not come back. One alternative is to set standard answers to explain the situation and that later the question will be answered when an expert increases the base.

Other problem may occur when wrong knowledge is stored. The advice is that only expert people and authorized users can use the administrative module. For that, the VA has a control module that validate users before accessing the administrative module. Furthermore, users can be registered as administrators or experts (or some kind to be created) and the privileges of each class may be set by super-users.

Finally, the VA technology may be used in many and different applications, being enough to change the knowledge base. The engine remains the same.

7.1 Future work

At the moment, the engine used in the VA only accepts knowledge codified in Portuguese. We are testing an engine to support English. The effort is minimum, since it is only necessary to change *stopwords* and the stemming algorithm.

The engine is being improved to analyze context of the questions. Currently, each question input by the user is considered independent of others. In some cases, it is necessary to analyze the context, that is, past questions. For example, if a user asks “*how much costs the product X*” and next asks “*and where can I buy it*”, the VA needs to understand that she/he is talking about the product X. This new feature is under test.

Soon, the VA will be using scripts for data collection. This will allow collecting opinions and demographic data through the VA, for example, to register users, to make marketing research or interview users.

We are also implementing a module for querying databases. This will allow presenting dynamic information to users, like products in stock, up-to-date prices and other data stored in tables.

In the same way, a future module will present news captured automatically from specialized websites. This module uses the technology presented in the paper of (SALDANÃ et al., 2003). Its main advantage is to enrich the VA with content about different subjects extracted automatically from the Web, without humans having to store knowledge.

Finally, the advanced version of the VA will work with sound, synthesizing voice for answers and recognizing audio questions.

8. REFERENCES

CHEN, Hsinchum (1994) The vocabulary problem in collaboration. **Computer**, 27(5), p.2-10, May.

CHOI, Wankyu; KENT, Allan; PRASAD, Ganesh; ULLMAN, Chris. (2001) **Beginning PHP4 Programando.** São Paulo: Makron Books, 719p.

DAVENPORT, T. H & PRUZAC, L. (1997) **Working knowledge** – how organizations manage what they know. Harvard Business School Press, 224 p.

NONAKA, I. & TAKEUCHI, T. (1995) **The knowledge-creating company**: how japanese companies create the dynamics of innovation. Oxford University Press, Cambridge, UK.

ORENGO, Viviane M & HUYCK, Christian. (2001) A Stemming Algorithm for the Portuguese Language. In: Symposium on String Processing and Information Retrieval, SPIRE'2001, Chile.

SALDAÑA, Ramiro et al. (2003) Captura automática de textos en la web para bibliotecas digitales. In: IV Coloquio Internacional de Ciencias de la

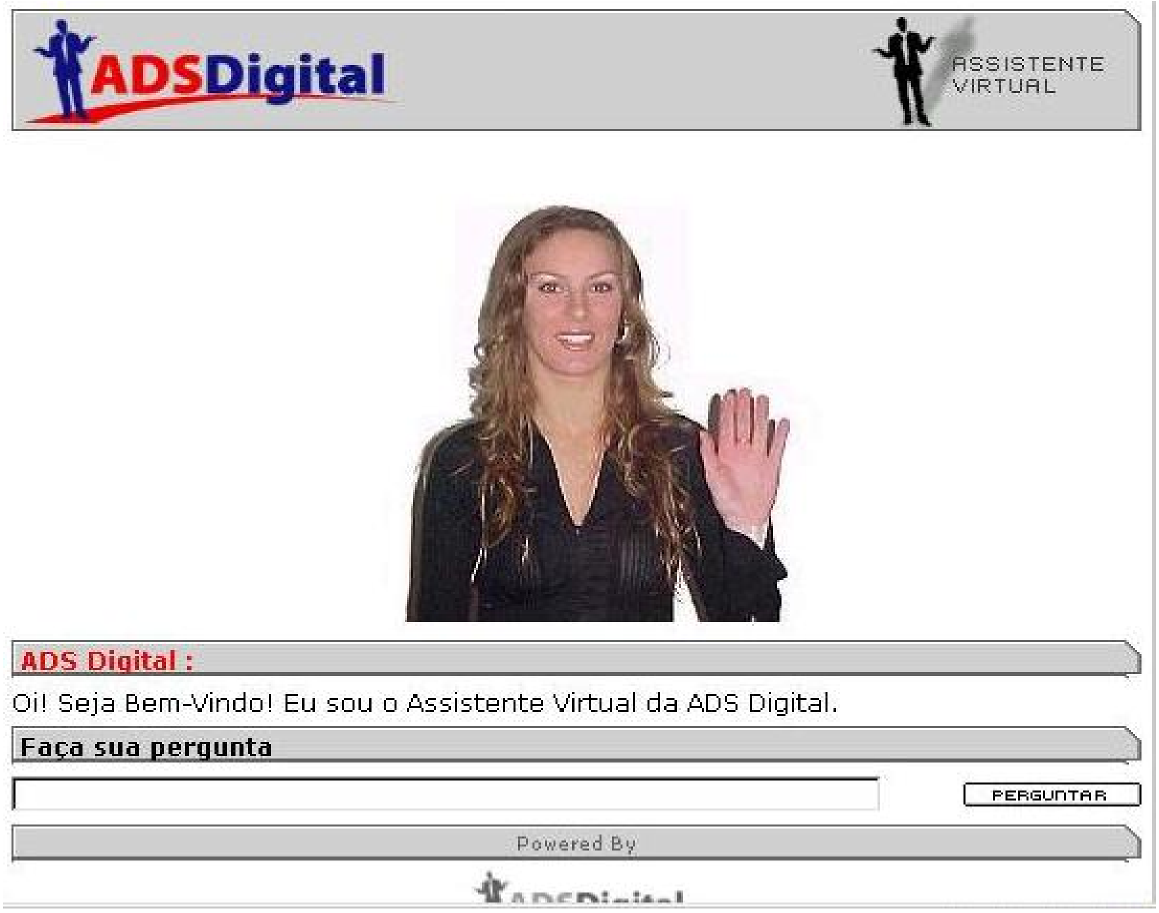
Documentación y VI Congreso del Capítulo Español de ISKO (International Society for Knowledge Organization), Salamanca, May.

SALTON, Gerard & McGILL, M. J. (1983) **Introduction to modern information retrieval**. New York: McGraw-Hill.

SENGE, Peter M. (2001) **The fifth discipline: the art and practice of the learning organization**. 9th edition. Best Seller, 444p.

WEIZENBAUM, J. (1967) ELIZA: a computer program for the study of natural language communication between man and machine. **Communications of the ACM**, v.10.

ZUE, Victor (1997) Conversational interfaces: advances and challenges. In: **Proceedings EUROSPEECH'97**. Rhodes, Grécia, p.9-18.



Oi! Seja Bem-Vindo! Eu sou o Assistente Virtual da ADS Digital.

Figure 1: A snapshot of a Virtual Assistant.