Road map for the implementation of a conversational agent chatbot consistent with the guidelines of the Design System Italy (DSI)

Davide Bruno

Regione Toscana, via di Novoli,27, Firenze, 50127, Italy

Abstract

The project idea intends to set some cornerstones for the design and subsequent implementation of a conversational agent based on intent and artificial intelligence generative with the aim of exploiting public service semantics (CPSV-AP_IT Core Public Service Vocabulary) to improve the interaction between citizens and the public sector (PS).

Keywords

generative artificial intelligence, public sector (PS), citizen , CPSV-AP_IT, CEUR-WS

1. Introduction

Artificial Intelligence Generative will certainly have a significant impact in the Public sector (PS). This technology offers opportunities to improve the efficiency, transparency and quality of public services. By implementing AI Generative-based systems, PS can automate repetitive processes, optimise data management and improve interaction with citizens. aims to improve productivity, accessibility and efficiency in service delivery. It leverages technologies such as machine learning, natural language processing (NLP) and natural language processing. These tools, known as 'conversational applications' or more commonly 'chatbots', exploit the ability of machines to understand natural language to handle requests and interactions with citizens in a timely and contextual manner. These applications will make it possible to improve the relationship with citizens, providing personalised and rapid responses, thus helping to optimise the delivery of public services. These highly innovative solutions will have to take into account the regulatory framework of reference for the PA and, in any case, avoid excessive dependence on suppliers that could quickly become technological lock-in.

Another risk that should not be underestimated at this time of feverish excitement on the subject is 'overpromising' and the illusion of perfection that comes from controlled demos where everything seems perfect and everything is solved simply by installing Artificial Intelligence Generative, at the moment there are not many solutions that are field-tested in real situations and are not known.

[1]*Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy*

✉ davide.bruno@regione.toscana.it ;

It is therefore crucial to choose the paradigm for the development of the target architecture and to identify generative AI models. It will be possible to exploit the great potential of generative AI if we take a cautious and forward-looking responsible approach
favouring open-source solutions or those based on shared open standards so as to guarantee PA flexibility and increase the possibility of changing suppliers in the future.
PAs will have to develop in-house skills very quickly by investing in staff training to make them at least aware and capable of expressing functional requirements, and not in the acquisition phase, but as mentioned, we must also be very vigilant in the operation and pre-operation phase, remembering that we are talking about new technological solutions that will certainly need to mature.
It will not be easy in this very dynamic and exponentially exploding field of LLM solutions, techniques and models. However, lasting collaboration with other public bodies, including research bodies, must be encouraged and institutionalised in order to share experiences and best practices.

2. Reference Models and Goals

The 'Core Public Service Vocabulary Application Profile' (CPSV-AP) is a data model designed to harmonise the description of public services on eGovernment portals. It provides a common vocabulary to describe public services, ensuring interoperability and standardisation between different implementations.[2]
The CPSV-AP aims at structuring public service information, making it user-centred and machine-readable, facilitating the creation of a public service catalogue that is interoperable and efficient. This vocabulary is used by several EU countries, including Italy, to describe public services and associated life events in a standardised way.

In 2017, the Agency for Digital Italy (AgID) published, in the OntoPiA controlled set of ontologies and vocabularies[3] (Ontologies for Public Administration), a special vocabulary for the definition of public services CPSV-AP_IT Core Public Service Vocabulary.[4]
The CPSV-AP_IT is the Italian version of this vocabulary, offering a framework to describe public services in Italy, in line with the European vocabulary of core public services. Designing and implementing a conversational agent natively integrated with CPSV-AP_IT will help improve the accessibility and effectiveness of public services, enabling users to interact in a more intuitive and personalised way with the information and services offered[5]. Moreover, the artificial intelligence of the conversational agent could be enhanced by the structure and semantics provided by the CPSV-AP_IT, facilitating a better understanding of user requests and offering more precise and contextualised answers

The realised artefact would be reusable at different administrative levels of the central and local Public Administration thanks to the work carried out over the years by AGID and by virtue of the fact that the Core Vocabulary of Public Services-Italian Profile (CPSV-AP_IT) has been defined. Moreover, the aspects of accessibility and usability would be guaranteed by including as a requirement by design the respect of the principles of accessibility and usability already present in the Design System Italia[6] (DSI).

3. Main steps for the design and implementation

The project idea therefore intends to set some cornerstones for the design and subsequent implementation of a conversational agent based on intent and generative artificial intelligence with the aim of exploiting the semantics of public services to improve the interaction between citizens and public administration.

The selection phase of a chatbot development paradigm/platform is crucial as already

[2]https://ec.europa.eu/isa2/solutions/core-public-service-vocabulary-application-profile-cpsv-ap_en/

[3]OntoPiA: GitHub: https://github.com/italia/daf-ontologie-vocabolari-controllati e GitHub wiki: https://github.com/italia/daf-ontologie-vocabolari-controllati/wiki (ita)
[4]Avaiable at link:<https://github.com/italia/daf-ontologie-vocabolari-controllati/blob/master/Ontologie/CPSV/v1.1/CPSV-AP_IT.rdf.>
[5]https://www.readygoone.it/approfondimenti/10-funzioni-dellassistente-conversazionale/
[6]https://designers.italia.it/design-system/

mentioned from a technological point of view it will be a mix of intent and generative AI.
For the reasons already expressed for the PA, open source solutions for generative AI should be favoured, combining it with retrieval-augmented generation (RAG) techniques[7].

Retrieval-augmented generation (RAG) has emerged as a promising solution that incorporates knowledge from external databases.
For knowledge systems, RAG has several advantages over the use of LLM alone:

Accuracy: RAG reduces and mitigates the risk of 'hallucinations', where LLMs might provide plausible but incorrect information. It does this by 'rooting' LLM answers in accurate data retrieved from your team's data sources to generate reliable answers.

Transparency: good RAG systems can provide references that allow users to verify where information comes from, adding a level of trust and accountability to the answers provided by RAG models.

Customisation: RAG systems can use data specific to your company or industry (e.g. naming conventions), making them adaptable and ensuring that answers are relevant to your specific context.

Other crucial elements in the PA's selection of the open source artificial intelligence architecture are:

Agnostic with respect to language models (it must therefore work with customised OpenAI, Cohere, HuggingFace models) The solution must possess long-term memory, have the possibility of using external tools (API, other models), be able to ingest documents of different formats (at least pdf, txt, json) and be developed with technologies that natively implement the possibility of scaling horizontally and vertically.

Non-functional requirements:
Accessibility by design: the chatbot should be implemented according to the accessibility by design paradigm while also taking into account the needs of user groups with disabilities as suggested by accessibility best practices1, such as the provision of text alternatives for any images and audio transcripts.
Multilingualism: implementation of multilingual functionality.

[7]https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/

3.1. Acquisition and preparation of data

1. **Retrieval of CPSV-AP_IT data**: Data in CPSV-AP_IT format (e.g. eligibility requirements, tariffs) and identification of relations with other services.
2. **Pre-processing and organisation**: In the case of data not conforming to CPSV-AP_IT, a mapping phase is still necessary to clean and harmonise the data in a format suitable for the chosen development architecture. It may be necessary to convert them into a computer-readable format (e.g. JSON, CSV) and to structure and optimise them for efficient queries.
3. **Identification of intentions and entities**: in this phase, the potential questions and intentions of the user (e.g. "how do I apply for a passport renewal?") and the entities they might mention (e.g. "passport", "renewal") should be defined. This could help the chatbot understand the user's needs.

3.2. Chatbot technological development

1. **Design the flow of the conversation**: Create natural language dialogues that guide the user in the discovery of services. Consider common user questions and create branching paths based on their answers. Prioritise clarity and conciseness and non-bureaucratic language.
2. **Integration of data with the CPSV-AP_IT model**: Linking the chatbot via ingestion pipelines to the previously processed CPSV-AP_IT data.
3. **Implementation Natural Language Processing (NLP)**: Using NLP techniques (intent recognition, entity extraction) to understand user queries and map them to relevant data points in the CPSV-AP_IT knowledge base. In order to enable the chatbot to retrieve accurate information about services.
4. **Error handling and fallback**: Implement mechanisms to handle user input that does not match defined intent or entity. Provide helpful hints or propose to rephrase the question. Consider offering a fallback option such as connecting with a human agent for

complex questions by retrieving what was typed.

3.3. Test and distribution

1. **Extensive testing**: Rigorous testing of the chatbot's functionality with various scenarios and user queries should be envisaged. Ensure that it accurately retrieves and presents information on public services, understands intent and provides clear guidance. Usability tests should also be envisaged.
2. **Monitoring and improvement**: Constant monitoring of the chatbot's performance. One should plan to collect feedback from users especially transactions that did not go well and use it to refine conversation flow, NLP accuracy and the overall user experience.

4. Conclusions and on-going activities

This presentation briefly describes a possible road map for the implementation of a conversational agent based on intent and generative artificial intelligence (AIG) with the aim of exploiting public service semantics (CPSV-AP_IT Core Public Service Vocabulary) to improve the interaction between citizens and the public administration (PA). Retrieval Augmented Generation (RAG) in general offers several advantages, in particular to improve the capabilities of artificial intelligence systems. In short, it is an approach that combines large language models (LLM) with information retrieval (IR) to improve the accuracy and relevance of LLM-generated text.

In a nutshell, the aims of this proposal:

1. Improving the discovery of online and on-site services of the public administration
2. Providing personalised and relevant answers to users
3. Reduce first level help desk calls to various services

An indicative road map of development:

- A prototype will be realised and validated by 2024.
- By first half of 2025 go live in production.

Possible future developments also automate the delivery of some simple services, integration at least as UX in the Design System Italy.

We may conclude by saying that the PA should not make the mistake of building an architecture that is bound to a single LLM model or specific solutions because depending on the specific use case, expected performance and costs, the configuration of the generative AI solution will be different.

Just to give an example of the variety and speed with which this sector is evolving, Anthropic alone released three LLM models between 2023 and 2024: #Claude1, #Claude2 and #Claude3.

OpenAI appears to be close to launching new versions of #GPT, which it claims will represent a further leap forward. The galaxy of generative artificial intelligence is still evolving strongly.
And it also has an impact on the open source world in fact the difference between an open-source LLM and close so binary is showing its limits or rather this new pradigma1 is establishing itself with respect to LLM so we can have the following types of models:

Openly Trained Models (OLMo, Pythia, etc.) - are those models with training data, training code and weights available without restrictions on use.
Permissible Usage Models (Llama**, Mistral, Gemma, etc.) - are those models with base model weights and inference code available for easy set-up and distribution.
Closed LLMs - everything from GPT4 to a random set of tuned weights without much information.

From the point of view of the Public Administration this is desirable a cautious and far-sighted responsible approach confirming the main requirement already expressed the framework selected must be agnostic with respect to the LLM model and in any case as PA we should prefer **Openly Trained Models** (OLMo, Pythia, etc.) or **Permissible Usage Models** (Llama**, Mistral, Gemma, etc.).

References

[1] Retrieval-Augmented Generation for Natural Language Understanding" di Patrick Lewis et al. (2020): https://arxiv.org/abs/2005.11401

[2] RAG: A Simple but Effective Approach to Neural Conversational Modeling" di Alexander Rush et al. (2020): https://medium.com/dropout-analytics/what-is-rag-in-generative-ai-f5b8c13575f8

[3] "RAG-BERT: Retrieval-Augmented Generation with BERT" di Honglei Zhuang et al. (2020): https://www.analyticsvidhya.com/blog/2023/10/rags-innovative-approach-to-unifying-retrieval-and-generation-in-nlp/

[4] "Towards Controllable and Consistent Generation with Retrieval-Augmented Generation" di Yilun Wang et al. (2021): https://aclanthology.org/2020.coling-main.207

[5] https://www.marktechpost.com/2024/04/01/evolution-of-rags-naive-rag-advanced-rag-and-modular-rag-architectures/Wang