iReason

s&t

Technical Architecture of OpenBrain - the Intelligent Voice Bot Framework
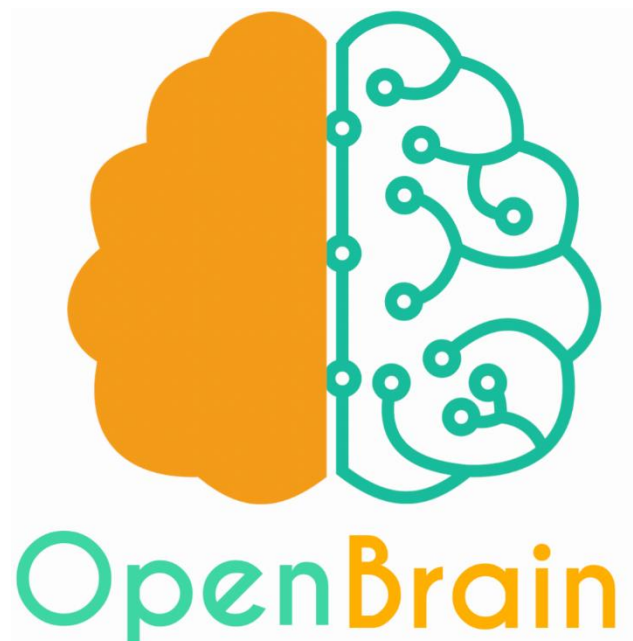
OpenBrain

Contents

1 Introduction

This White Paper provides a technical overview and some insights into the technologies that OpenBrain Voice Bot Framework relies on. The purpose is to help various service providers to understand the technical architecture of the OpenBrain Intelligent Voice Bot Framework.

1.1   How can OpenBrain be used?

OpenBrain is a modular framework comprised of intelligent units aiming to make the service providers more productive in both B2B and B2C communication. OpenBrain automatically unifies internal and external sources of knowledge, gathers and extracts meaningful information, then translates it into a human-like manner on user demand. Thus, the employees and the customers can efficiently provide and require information without moving from one information source to another or wasting additional human resources for technical support, as is the case in the existing IVR technologies. As a result, OpenBrain can easily transform the companies' IVR systems into intelligent IVR solution, which can reside both as an on-premise solution or a service in the Cloud.

Thanks to the wide range of functionality, OpenBrain can be applied in a variety of solution areas with specific target on:

- Business intelligence

- IVR systems intelligence

- Assistive technologies

OpenBrain provides structure and meaning to large volumes of unstructured content. It enables customers and employees to find and use the information in a human-like manner of communication (by asking and acquiring the response in spoken or written natural language). The framework is content-neutral and can integrate with any source. It can transform your web offer into a knowledge graph, upgrade your IVR system with intelligent speech and information retrieval technologies, and transform your web interface into accessible information in compliance with assistive technologies demands.

The unique advantage of OpenBrain technology is its capability to deploy all AI models, including text-to-speech, speech-to-text, dialog flow control, knowledge management, as an in-house solution without accessing third-party services and solutions. Hence, the transport of sensitive information with the company's internal services and databases is wholly secured and invisible to the external world. This concept prevents many security issues, including data leaks and data breaches, when using the complex set of technologies that any intelligent voice bot requires to operate.

1.2   Technical Standards

OpenBrain's backend is based on Deep learning methodologies, NLP Transformer architecture, and language representation methods based on Knowledge Graphs. It supports HTTP(S), RESTful communication, and WebSocket over encrypted TLS connection (WSS) protocol for streaming data and other asynchronous traffic. The interface is based on React, Spring Boot, and Python. OpenBrain operates on a Linux Server, and users only need a standard web browser. It supports Unicode UTF8 characters, which cover most languages.

OpenBrain scales to thousands of users. It supports load management, which is a method where simultaneous user requests are distributed and balanced among multiple servers. The intelligent models can run on CPUs, however, GPUs are preferred to speed-up the inference. The evaluation results are as follows.

- The CPU evaluation is performed on Intel(R) Xeon(R) CPU E7-8880 @ 2.30GHz
- The GPU evaluation is performed on the NVIDIA Tesla P100 device. All CUDA supporting GPU devices are applicable.
- Number of recommended users per server configuration providing real-time communication (1s audio is processed for less than 1s) with the virtual agent.
- The OpenBrain platform supports horizontal scaling to infinite number of nodes.

	Concurrent Users
CPU (8 cores) + 12GB RAM	5x
GPU (Tesla P100) + 12RAM	58x

OPENBRAIN CONCURRENT USERS
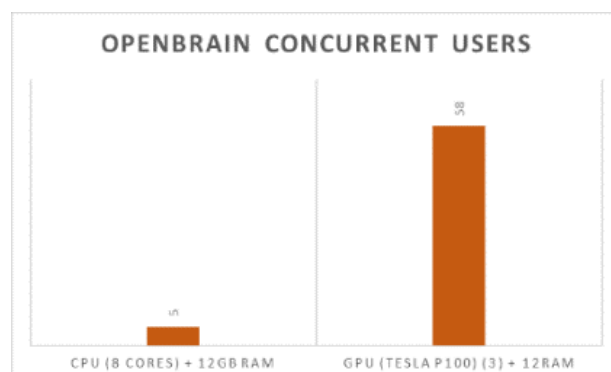
1.4   Relations to other types of voice bot frameworks

OpenBrain framework is modular and is specially designed to help businesses manage inner knowledge and provide services to customers in a human-like manner. The framework includes a wide range of unique functionalities not presented in other standard voice bot frameworks offered by other competitors as none of them offers on-premise solution. Most importantly, our framework supports Balkan languages using specially tailored ontologies, algorithms, and models trained on a large corpus of texts written in Balkan languages.

The following table presents 42 companies that provide chatbot platforms around the globe. Only 8 of them provide NLP engines (highlighted with orange) and even more, only 4 provide voice and text services (bolded) and thus can be considered as competitors, whereas the others are not considered intelligent chatbots. All 4 companies are from the USA, meaning there is no visible competitor in the country, in the Balkan, or the wider European area.

Company/Bot Name	Location	Founded	Capabilities	UX	Function	References	Languages
CM.com	Netherland	2005	Chatbot Platform	GUI	Voice / Text	Eneco, ANWB, Missguided, A.S.R	Multilingual
FlowXO	UK	2014	Chatbot Platform	GUI	Text	N/A	Multilingual
Sor'un	Turkey	2014	Chatbot Platform	GUI	Text	Garanti, Hepsiburada, Pizza Hut, Decathlon	Turkish
Rasa NLU	Germany	2015	**NLP Engine**	Programmable	Text	UBS, Raiffeisen, Yellow Pages, Lemonade, Adobe	English, German
Recast.AI (acquired by SAP)	France	2015	Chatbot Platform	Programmable	Text	Accenture, Capgemini, Deloitte, SalesForce, Cisco, Facebook, Slack	Multilingual
Twyla	Germany	2015	Chatbot Platform	GUI	Text	Shopify, Salesforce, Oracle	English
Cognigy	Germany	2016	Chatbot Platform	GUI	Voice / Text	Henkel, T√úV Rheinland, Salzburg AG	Multilingual
Destygo	France	2016	Chatbot Platform	GUI	Text	Iberia, SNCF, RATP, Kilroy	Multilingual

Engati	India	2017	Chatbot Platform	Program mable	Text	IKEA, Pwc, Amadeus	Multilingual
FloatBot	India	2017	Chatbot Platform	Program mable	Text	Andhra bank, Lifecell, Pimpri-Chinch wad	Multilingual
Morph.ai	India	2016	Chatbot Platform	Program mable	Text	Manchester City, Yamaha, Yes Bank	Multilingual
Amplify.ai	USA	2016	Chatbot Platform	GUI	Text	Change.org	Multilingual
Beep Boop	USA	2015	Chatbot Platform	GUI	Text	N/A	Multilingual
Bottr	USA	2015	Chatbot Platform	GUI	Text	N/A	Multilingual
Chatfuel	USA	2015	Chatbot Platform	GUI	Text	Adidas, Buzzfeed, MTV, Techcrunch, Volkswagen, Uber	Multilingual
Conversable	USA	2016	Chatbot Platform	GUI	Text	Budweiser, EY, PG, Pizza Hut, Sony, TGI Fridays, Viacom, Whole Foods,	English
Google Dialogflow (previously Api.ai)	**USA**	**2010**	**NLP Engine**	**Program mable**	**Voice / Text**	**WSJ, KLM, Giorgio Armani, Comcast**	**Multilingual**
Gubshup	USA	2014	Chatbot Platform	GUI	Text	Facebook, Linkedin, Citibank, Twitter, Venture Beat, Cisco, Unilever, Ebay	English, Hindi
IBM Watson Conversation Service	**USA**	**2014**	**NLP Engine**	**Program mable**	**Voice / Text**	**N/A**	**English, Japanese**
KITT.AI (acquired by Baidu)	USA	2014	Chatbot Platform	GUI	Text	N/A	English
ManyChat	USA	2015	Chatbot Platform	GUI	Text	Neil Patel, School of Bots, Rapid Fire Pizza	English
Microsoft Bot Framework	USA	N/A	**NLP Engine**	Program mable	Voice / Text	N/A	Multilingual
Microsoft Language Understanding Intelligent Service (LUIS)	**USA**	**N/A**	**NLP Engine**	**Program mable**	**Voice / Text**	**N/A**	**Multilingual**
Mindmeld (acquired by Cisco)	USA	2011	Chatbot Platform	Program mable	Voice / Text	N/A	English

motion.ai (acquired by HubSpot)	USA	2015	Chatbot Platform	GUI	Text	N/A	Multilingual
Octane.ai	USA	2016	Chatbot Platform	GUI	Text	Kiehls, Warner Music Group	English
Pandorabots	USA	2008	Chatbot Platform	Programmable	Text	N/A	Multilingual
Pypestream	USA	2015	Chatbot Platform	Programmable	Text	Cisco, Microsoft, Salesforce	English
Recime	USA	2017	Chatbot Platform	GUI	Text	Coca Cola, Wyre, Gigigo	Multilingual
Reply.ai	USA	2016	Chatbot Platform	Programmable	Text	Mapfre, Samsung, Starbucks, Hilton	English
Semantic Machines	USA	2014	**NLP Engine**	Programmable	**Text**	N/A	**Multilingual**
wit.ai (acquired by Facebook)	**USA**	**2013**	**NLP Engine**	**Programmable**	**Voice / Text**	**N/A**	**Multilingual**
Yekaliva.ai	USA	N/A	Chatbot Platform	GUI	Text	N/A	English
Meya.ai	Canada	2015	Chatbot Platform	Programmable	Text	Sony, Vodafone, Mercedes Benz	Multilingual
Code-well	**Macedonia**	**2011**	**Chatbot Platform**	**N/A**	**Text**	**A1**	**MK**
iReason.mk	**Macedonia**	**2021**	**NLP Engine**	**Programmable**	**Voice / Text**	**N/A**	**Multilingual + MK**

All the platforms that can be considered as competitors are commercially available, thus we have evaluated the functionalities offered as either with N/A (not available) or A (available). From the following table, it can be perceived that iReason's OpenBrain is the only intelligent voice bot framework available in all Balkan languages that offers its services On-premise, thus assuring all the company's information stays in-house.

Product analysis	iReason OpenBrain	Google Dialogflow	IBM Watson	Microsoft LUIS	Wit.ai (Facebook)	Code-well Slavcho
Natural Language Understanding	A	A	A	A	A	N/A
Natural Language Generation	A	A	A	A	A	N/A
Named entity recognition	A	A	A	A	A	N/A
Intent detection	A	A	A	A	A	A
Task classification	A	N/A	A	A	A	A
Speech-to-text	A	A	A	A	A	N/A
Text-to-speech	A	A	A	A	A	N/A
Contextual conversation	A	A	A	A	A	A
Sentiment Analysis	A	A	A	N/A	N/A	A
Opinion mining	A	N/A	N/A	N/A	N/A	A
Question answering	A	A	A	A	N/A	A
General knowledge	A	N/A	N/A	Maybe	N/A	N/A
Summarization	A	N/A	A	N/A	N/A	N/A
Multilingual	A	A	A	A	A	A
Graph knowledge	A	N/A	N/A	N/A	N/A	N/A
Dialog-flow management system	A	A	A	A	A	N/A
On-premise	A	N/A	N/A	N/A	N/A	N/A
SaaS	A	A	A	A	A	A

2 The Layers of OpenBrain

The architecture of OpenBrain consists of five layers:

1. **Knowledge Layer:** at the core of OpenBrain are the knowledge graphs and summarization models. This layer represents the inner knowledge and the general knowledge on demand.

2. **NLP Layer:** encompass the multilingual natural language understanding and generation engines, as well as the specific models for named entity recognition, intent and task classification, opinion mining, sentiment analysis and question answering models.

3. **Application Layer:** at this layer reside the text-to-speech and speech-to-text intelligent models, typo correction models, and also the dialog-flow management system for creating novel custom-based scenarios.

4. **Security Layer:** controls access to OpenBrain, authenticates and authorizes users.

5. **User Interface:** is the final layer that meets the end-user of OpenBrain and is responsible for assuring contextual voice, text, or combined conversation.
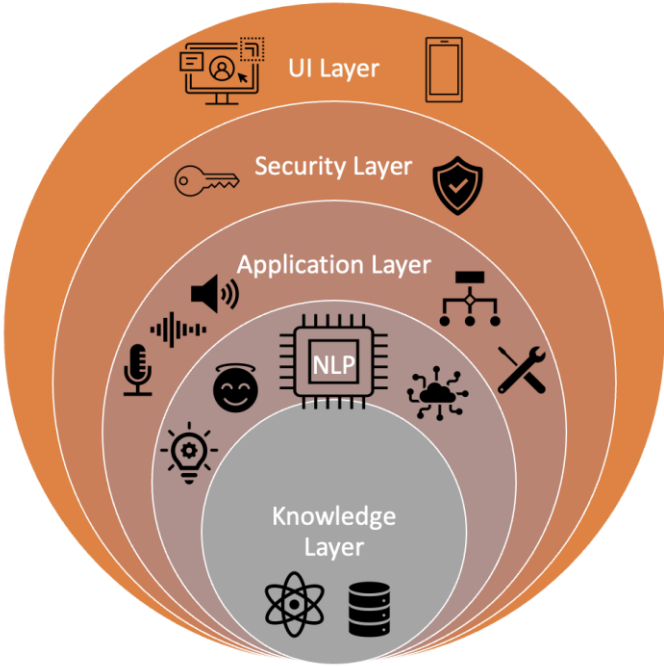
Each of these layers will be described further in the following subsections.

The NLP is the heart of the framework that pulls the intelligent models at each of the layers, enabling human-like communication as shown in the following Figure.
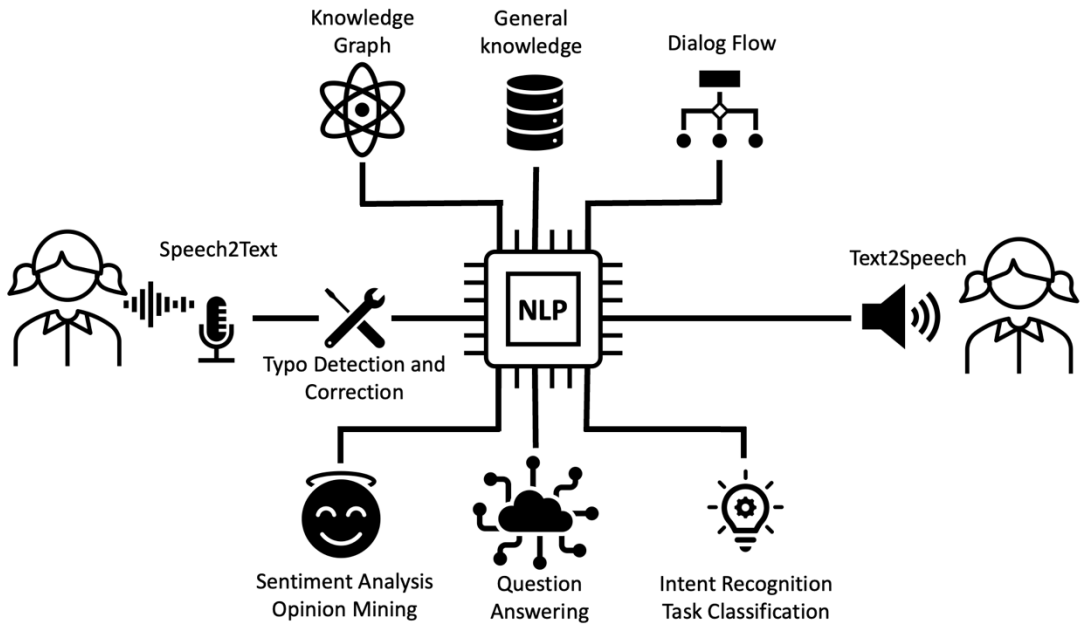
2.1 Knowledge Layer

Knowledge is the core of the OpenBrain framework and crucial for the efficiency of the instances produced from the framework. Knowledge can be retrieved both from internal sources (the inner databases of the company) or external sources (such as the Internet, Wikipedia, Dbpedia, etc). Data is usually obtained in semi-structured format and needs to be additionally processed to enable the extraction of meaningful knowledge. Knowledge graphs approach is used to organize the data in terms of related entities and their attributes, thus the search for information is fast and accurate.

2.2 NLP Layer

NLP layer is crucial to managing the knowledge as structured in the Knowledge Layer. This layer consists of many intelligent models built upon the knowledge and the natural language of interest for the particular instance of OpenBrain, since this framework is multilingual, meaning it can be easily adapted to any language for which there are available resources. The framework offers trained models for natural language understanding and natural language generation as most important feature that assures the voice bot clearly understands the morphological characteristics of a language, and thus can understand the user's requests and is able to appropriately respond to those requests in terms of serving the demanded information, or, even more, to trigger suitable actions.

Communication is a complex process for which understanding a language is not the only problem that should be considered. Human beings have emotions, ask complex questions, ask for help, and usually intent to take some actions when communicate with intelligent voice bot in both web and an IVR environment. For those reasons, NLP layer holds few more intelligent models built upon Deep learning methodologies:

- Sentiment analysis and opinion mining to understand the user's emotions and to be able to smoothen an intention for rude conversation.
- Question answering intelligent module that is able to search for answers from complex questions given the knowledge graph in the Knowledge Layer.
- Intent recognition and consequently task classification modules that are trained on a particular domain of interest, able to extract the focus and the entities of the conversation and to recognize the intent of the user.

2.3 Application Layer

The Application layer holds few more intelligent models based on Deep learning methodologies that are crucial for the human-like communication experience, and those are:

- **Text-to-speech engine** – Deep learning based model built upon large corpus of single-speaker audio files able to synthesize human-like speech for a particular language. The model can be trained on any language in a short time.

- **Speech-to-text engine** - Deep learning based model built upon hundreds on hours of speech from hundreds of speakers, transcribed in textual form. The model can be trained on any language within few months.

- **Intelligent typo correction model** – This model works in pair with the speech-to-text engines with the aim to correct the misspelled words.

- **Dialog flow management system (DFMS)** – This GUI is a unique possibility for the company to build scenarios on its own. This GUI is made only for the service providers with the aim to ease the process of upgrading the voice bot instance with new intents, scenarios and conversation flows. The following figure describes a case of voice bot usage in the banking sector, in which the administrators can create their custom blocks and define transitions among multiple blocks to complete one use case. At its background, a code is automatically generated without any human intervention, and it is fed as an input to the complex pipeline of language models. Furthermore, DFMS provides the use of specially designed secured objects named Webhooks that allow efficient and scalable integration and interoperability with the company's internal services and databases.
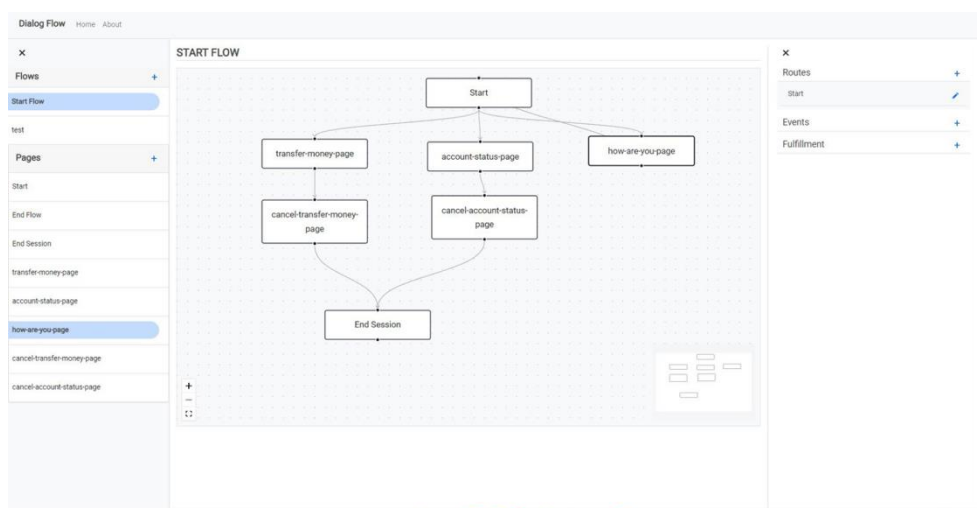

Dialog flow management system.

2.4 Security Layer

The content is secured by few key services that allow:

- Authentication – Since the administrators of the system are the only users that need to log in to the system, Single Sign-On (SSO) is the software authentication method which enables a user to log in once and gain access to the resources of multiple subsystems.

- Access Control - Additional functionality for stopping automated attempts to gain unauthorized access to content.

- Logging - Logging of events. Data modifications and Administrative actions are logged.

- Encryption - Key system information and passwords are encrypted.

- VPN - Traffic between the Server and Client can be encrypted using SSL or restricted through VPN.
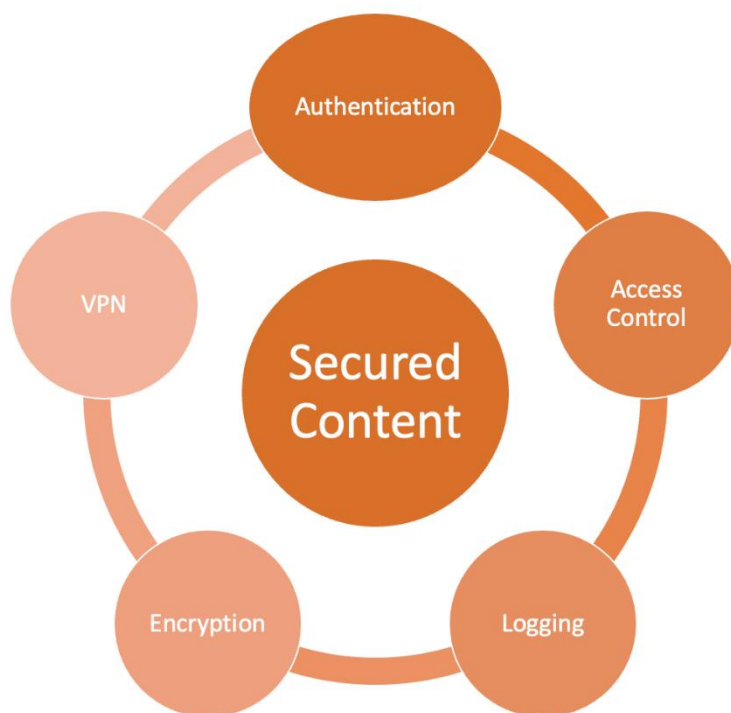
Security Layer of Open Brain Framework

2.5 User Interface Layer

The User Interface Layer offers three possible implementations and each of them is described as follows:
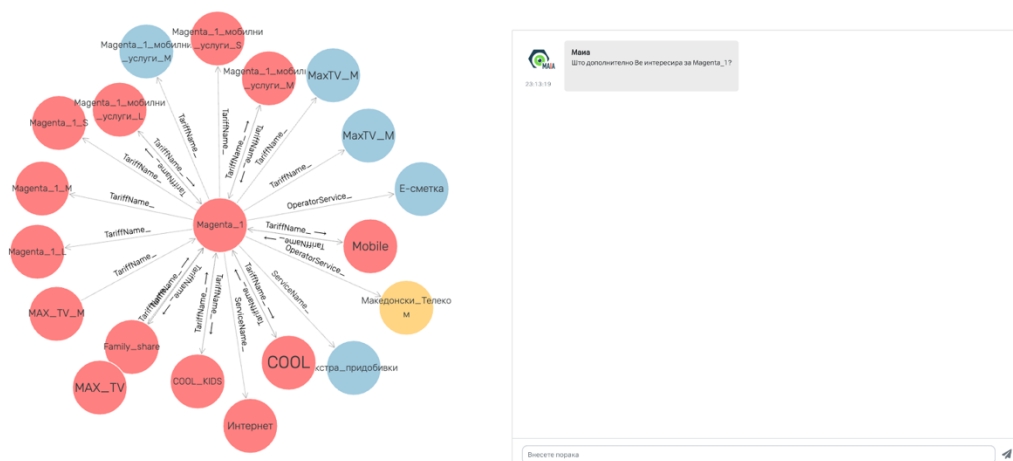
- Web view – This view is a simple user interface with chat environment and options to use the speech-to-text functionality to interact with the voice bot by using a speech, and an option to activate the text-to-speech functionality and obtain the answers in audio format. This view has also an admin panel, allowing the administrators to update the knowledge database with unstructured text, ready to be preprocessed by the models in the Knowledge Layer.

  The web view offers the possibility of being customized depending on the modules chosen to be active when creating a voice bot instance from the OpenBrain framework. The simple view can be upgraded with a view at the Knowledge Layer, presenting the knowledge graph and the active nodes at the moments a particular user is searching for information.

  The following figure presents examples from both simple web view and web view with knowledge graph activated from the OpenBrain's voice bot instance in Macedonian language (MAIA) for a banking and telecommunication sector, correspondingly. MAIA is further described in the following chapter.
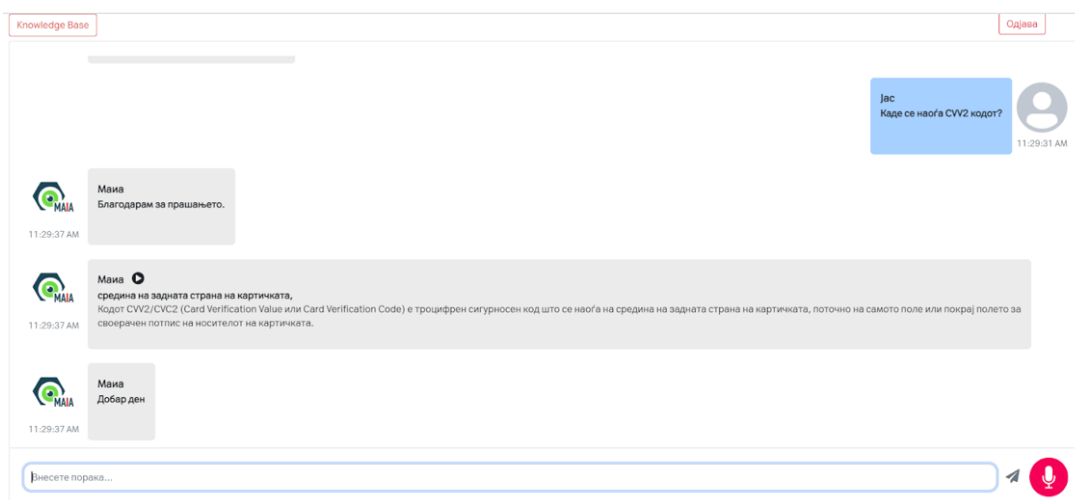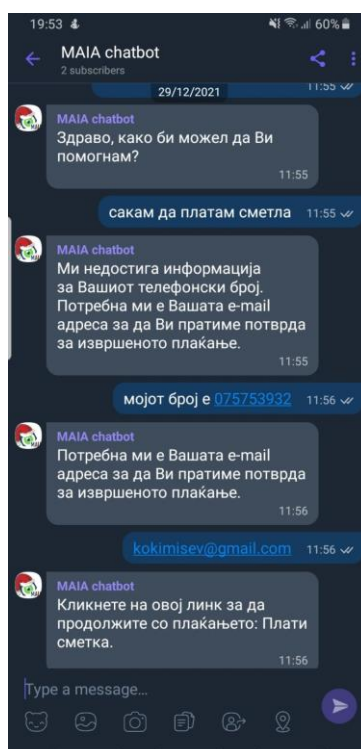

Simple web view with text-to-speech and speech-to-text engines activated.

Web view with activated knowledge graphs.

- Mobile phone view - This user interface is an integration with some popular apps such as Viber, Messenger, and Whatsapp. The aim of this view is to ease the communication with the voice bot by using a smart phone. An example of this view is provided in the following figure.

An example of Viber integration.

- IVR view - The user interface, in the case OpenBrain's instance is integrated with IVR system, is accessed via cellular line by the user. The text-to-speech and speech-to-text engines are active during the whole process, together with the other intelligent models from the Application, NLP and Knowledge Layer. Since this interface is not visible as the previously presented user interfaces, the integration with an IVR system is depicted in the following figure.

How a call moves through the IVR system with MAIA
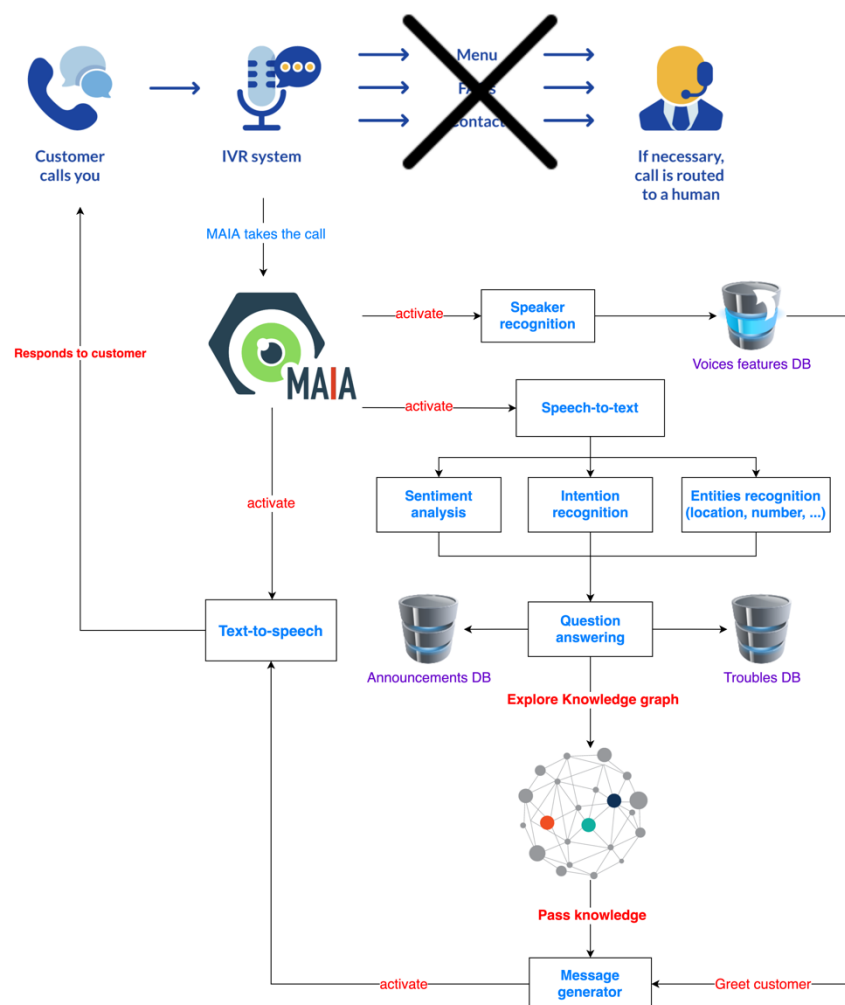
The intelligent voice bot takes the call and activates all the intelligent modules to solve the user's problem or request. If the intelligent voice bot is not able to solve the demands, the call is routed to a human resource as it is done in a traditional IVR system.

3 MAIA – an Instance of OpenBrain Framework

MAIA is an example of an instance from OpenBrain framework and presents the first Macedonian intelligent voice bot. MAIA integrates the first text-to-speech in the Macedonian language, the first speech-to-text in the Macedonian language, along with all the AI modules available at the different OpenBrain layers.

MAIA aims to test the OpenBrain framework in various fields, including the finance/banking sector and telecommunications. The interfaces previously shown are examples of multiple modules activations. The evaluation of MAIA has been demonstrated that MAIA is able to:

- Ease human resources
- Constantly learn and fine-tune the intelligent models as new data is available
- Can reside On-premise and in the Cloud
- Enable the creation of new dialog flows and conversation management via Dialog flow management system
- More efficient information retrieval in the Macedonian language

4 Business Model

OpenBrain framework is modular, meaning a voice bot instance can be composed on demand. Basic functionalities to check the quality of text-to-speech and speech-to-text services are offered in a demo version for free, along with administrator interface that allows the insertion of new data, and Question Answering module that allows a user to ask questions and obtain answers. Free demo can be acquired by completing the form at www.ireason.mk.

Further activation of modules can be acquired by requesting a Proof-Of-Concept (PoC). A PoC is available only as a service that resides in the Cloud, with a limited Question Answering and Intents capacities – 250 questions and 250 domain specific intents to be recognized, correspondingly.

A voice bot instance ready for production and public launch is a fully functional product based on specific requests by the company that requires a voice bot integration. The voice bot instance can be set as a service in the Cloud, or, On-premise depending on the company's preferences. There are two licensing options available, perpetual and subscription per year.

For further information contact us at contact@ireason.mk, or by completing the form available at www.ireason.mk.

5 About iReason and S&T collaboration

iReason is a company that aims to translate the research into original products. It is founded by researchers with academic backgrounds and Ph.D. in computer science and data science. OpenBrain is the first of the kind that can be offered On-premise and competes with giants such as Google, IBM, and Amazon. iReason's mission is to create original intelligent products according to users' specific needs. The company was founded in 2021 and is based in Skopje, N. Macedonia, with global representation through an extensive partner network of S&T.

More about the publications and the product developed so far can be read at www.ireason.mk.

iReason holds all Intellectual Property Rights with regards to all software the company has developed.

For more information, please contact us:
Phone: +389-75-753-932
E-mail: contact@ireason.mk
Web: www.ireason.mk

_____