NPFL123 Dialogue Systems

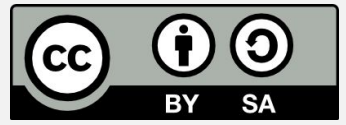
9. Dialog Authoring Tools

https://ufal.cz/npfl123

Ondřej Dušek, Mateusz Lango, Ondřej Plátek & **Jan Cuřín**

17. 4. 2025

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics



Short Intro of Jan Cuřín

Education

- MFF UK – PhD in 2006 at IFAL (Statistical Machine Translation)
- Johns Hopkins University, Baltimore, MD - 1999 NLP workshop at CLSP (MT: EGYPT -> GIZA)

Work Experience

- MAMA AI, Telma AI – 2021- Co-Founder, CTO
  - Natural Language Processing, Conversational Agents, AI
- IBM – 2004-2021 – Research Scientist, Manager at IBM Watson R&D Lab
  - Conversational Systems, NLP Technologies
- IFAL, MFF UK – 2002-2004 – Researcher, PhD Student
  - Machine Translation, NLP Technologies
- Schemantix – 2000-2001 – Software Engineer
  - Machine Translation, XML Technologies
- IFAL, MFF UK – 1998-2000 – Researcher, MSc. Student
  - Machine Translation, Parallel corpora for Prague Dependency Treebank (PCEDT)
- Intax – 1996-1998 – Member of translation/localization team
  - Localization of apps and operating systems for IBM Czech Republic

IBM Watson Services
- Watson Assistant
- Watson Speech (STT/TTS)
- Watson Language Translator

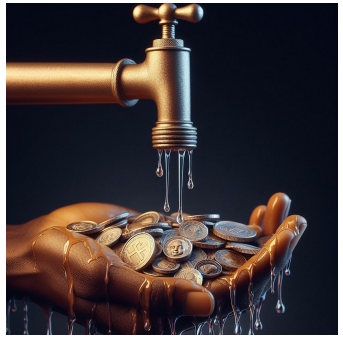
Business Use Of Dialogue Systems

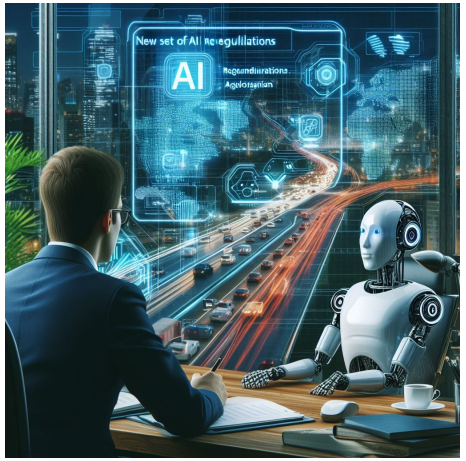
- Chat bots on web pages
  - Navigation through the content of the web pages (smart search)
  - Frequently asked question (FAQ), Product search, Product recommendation
  - Customer support
- Mobile applications with open text input
  - Domain specific apps with chatting functionality, banking apps
  - ChatGPT, ChatGPT+, BING, Copilot, …
- Assisting systems
  - Intra company “expert” system, Support for human operators
  - Data via chat (chat with your data sources), Automated reporting (chat with your data history)
- Speech based systems
  - Call centers automation – handling top x% of traffic
  - Outbound calls (to inform or collect feedback, incident resolution)
- Automotive applications
  - Search, calls, navigation, infotainment/entertainment (music, POIs)
- Infotainment systems and hospotiality
  - Infotainment systems for hotels, banks’ lobbies, home, etc …
  - Marketing campaigns, E-mail/WhatsApp channels (FAQ, quotes,)

Challenges of Creating Dialog Systems

- Data
  - Use of existing of human-to-human conversations
  - Use of intracompany structured and unstructured data
  - No data, just ideas
  - No idea at all
  - “just” generic LLMs use
- Scenarios, use-cases
  - Single domain
  - Single domain with chit-chat capability
  - Multi-domain

Generative AI


business



AI Regulation

Authoring a Dialogue System “old style”

Restaurant booking scenario

- System: Hello, this is Chez Pépé restaurant reservation system. How may I help you?
- User: I would like to book a table for tomorrow for 5 people
- System: From what time?
- User: From 7pm
- System: OK, I will make a reservation for tomorrow at 7pm, table for 5.
- User: Great!
- System: Looking forward to see you soon.

Authoring a Dialogue System “old style”

Restaurant booking scenario

Intents

#reserve_table
- I would like to book a table for 5
- Can I make a reservation?
- I need a reservation for tomorrow

#opening_hours
- Until when are you open?
- What are the opening hours?

#cancel_reservation
- I made my reservation yesterday, I want to cancel it.
- We could not make it today, may I cancel the reservation?

Entities

I need reservation for 5 people for tomorrow at 7pm

Next Friday from 6pm

For two.

@date
@time
@number
@restaurant_location

Dialogue

Welcome
Book a table

Entity	Context variable	Req.
@date	$res_date	Y
@time	$res_time	Y
@number	$guests	Y

Opening hours
Cancel reservation
- Yes
- No

<default answer>

Intents

- Collection of example how users will trigger the intent
- Usually corresponds to the actions supported by the dialog

- Intent model can be trained even on a small set of examples
- Word and sentence embeddings, stemmer, lemmatizer
- Bigger data collection needed for production system
- Data augmentation by large-language models (LLMs)

- Ordered n-best lists with confidences
- Use of intent n-bests in the dialog – disambiguation

Intents

#reserve_table
I would like to reserve a table for 5
Can I make a reservation?
I need a reservation for tomorrow

#opening_hours
Until when are you open?
What are the opening hours?

#cancel_reservation
I made my reservation yesterday,
I want to cancel it.
We could not make it today,
may I cancel the reservation?

Entities

~ Named entities recognition (NER)

- Different type of entities
  - Prebuilt (system) entities
    - Numbers, dates, time, GEO location, person names, units, currency
  - Domain catalogues
  - User defined entities
    - Gazetteers – fixed list of entities/synonyms
    - Regular expression based
    - Sequence labelling model based on sample annotations (contextual entities)

Entities

I need reservation for 5 people for tomorrow at 7pm

Next Friday from 6pm

For two.

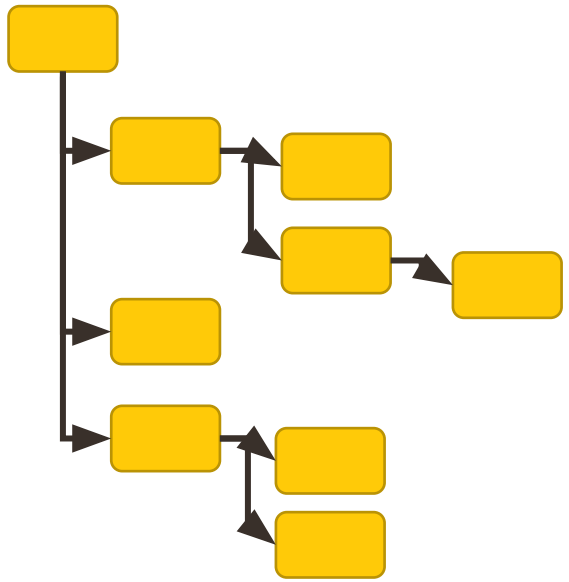
@date
@time
@number
@restaurant_location

Dialogue Flow/Tree

- Slot filling style (linear dialog)
  - Set of slots to fill is (required/optional)
  - Able to fill all slots partially or at once
  - Asking just for missing information
  - Ability to customize questions and answer for a particular slot
  - Ability to correct already filled information
  - Tight to user variables

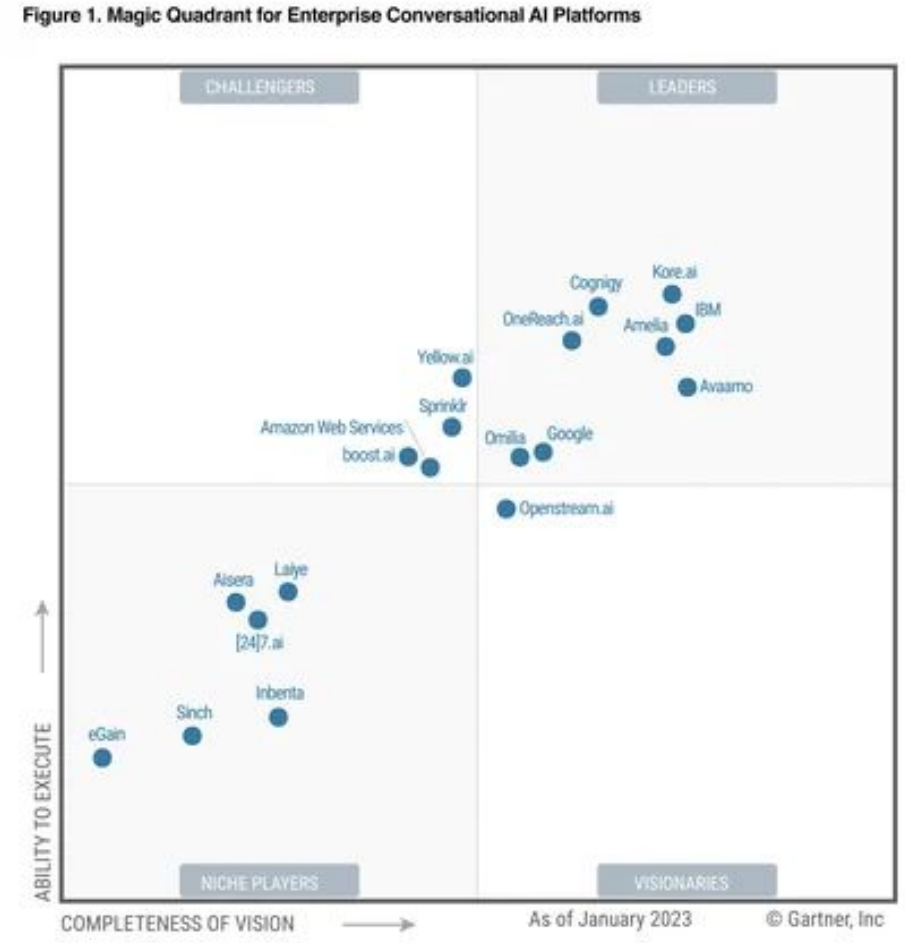
Entity	Context variable	Req.
@date	$res_date	Y
@time	$res_time	Y
@number	$guests	Y

- Dialogues tree (non-linear dialog)
  - Dialogue flow driven by a tree or graph structure
  - Conditions to get to the individual nodes of the tree/graph
  - Fallback strategies (none of the conditions is specified)



Gartner Magic Quadrant for Enterprise Conversational AI Platforms 2023

- Evaluation of conversational AI platforms in 2023
- IBM is historically performing well, still among the best in the Completeness of vision



Source: article in [CXToday]

Authoring tools

- IBM watsonx Assistant
  - https://www.ibm.com/cloud/watson-assistant/
  - Video tutorial: https://console.bluemix.net/docs/services/assistant/tool-overview.html
- Google Dialog Flow
  - https://dialogflow.com/
  - Video tutorials: https://cloud.google.com/dialogflow/docs/video
- Amazon Alexa Skills
  - https://developer.amazon.com/alexa-skills-kit
  - Video tutorial: https://www.alphavoice.io/video/alexa-developers/alexa-skills-kit-developer-console-build
- Microsoft Cortana Skills
  - https://developer.microsoft.com/en-us/cortana
- Apple SiriKit (Siri-enabled iOS apps)
  - https://developer.apple.com/sirikit/

Demo

- Sample chatbot in IBM watsonx Assistant

Restaurant booking scenario

http://www.bienvenuechezpepe.com/

Features/Concepts

- Dialogue context / history
  - Condition on context variables collected in previous turns
  - Reference/anaphora resolution using collected variables
- Fallback strategies / Digression
  - Allow "jumping" to different topic for a while and then return back
- Disambiguation support
  - Similar confidence of multiple choices – ask user to select
- Calling external APIs
  - Webhooks/Cloud functions …

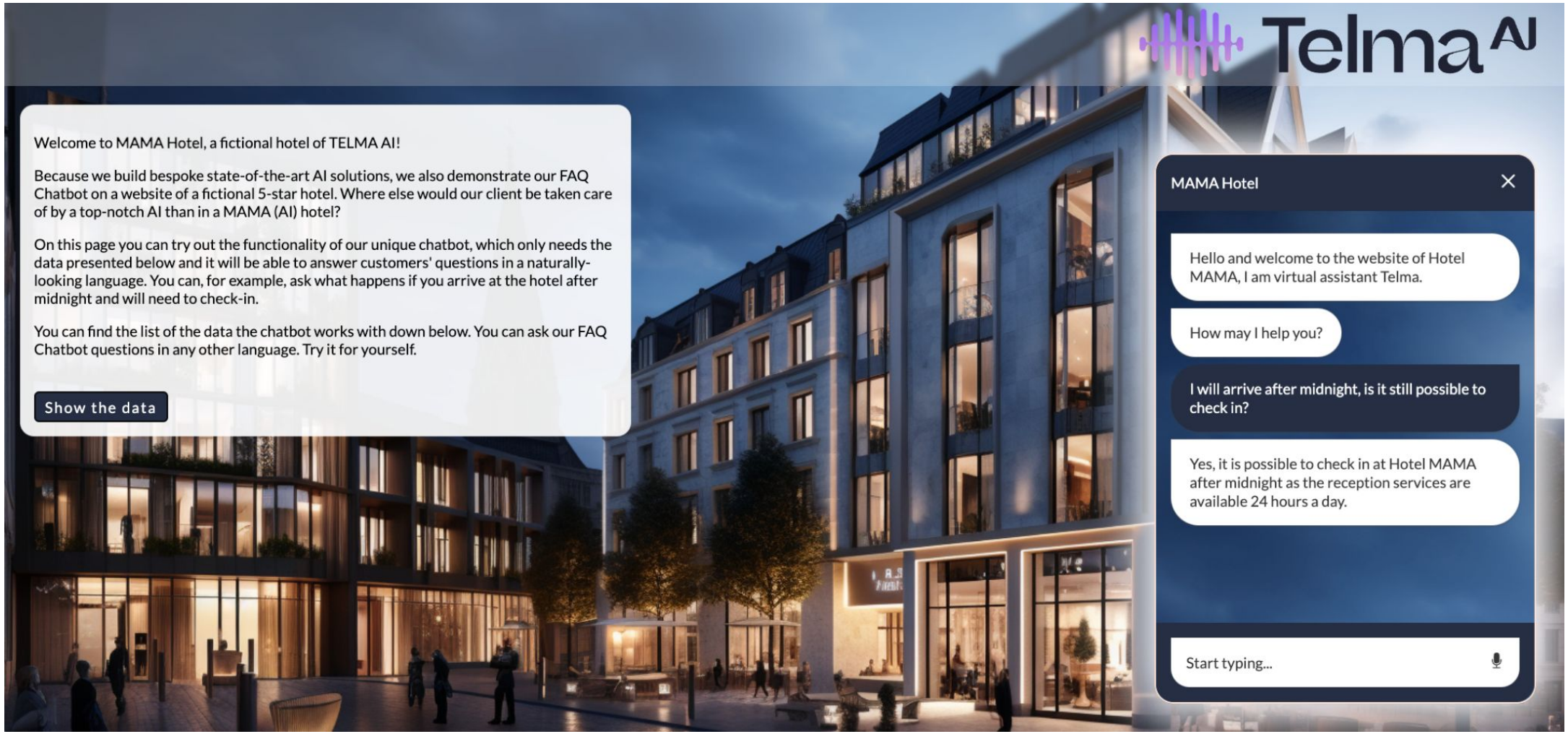
~ Use of LLMs for “natural” response generation and chitchat

Maintaining and improving chatbot in production

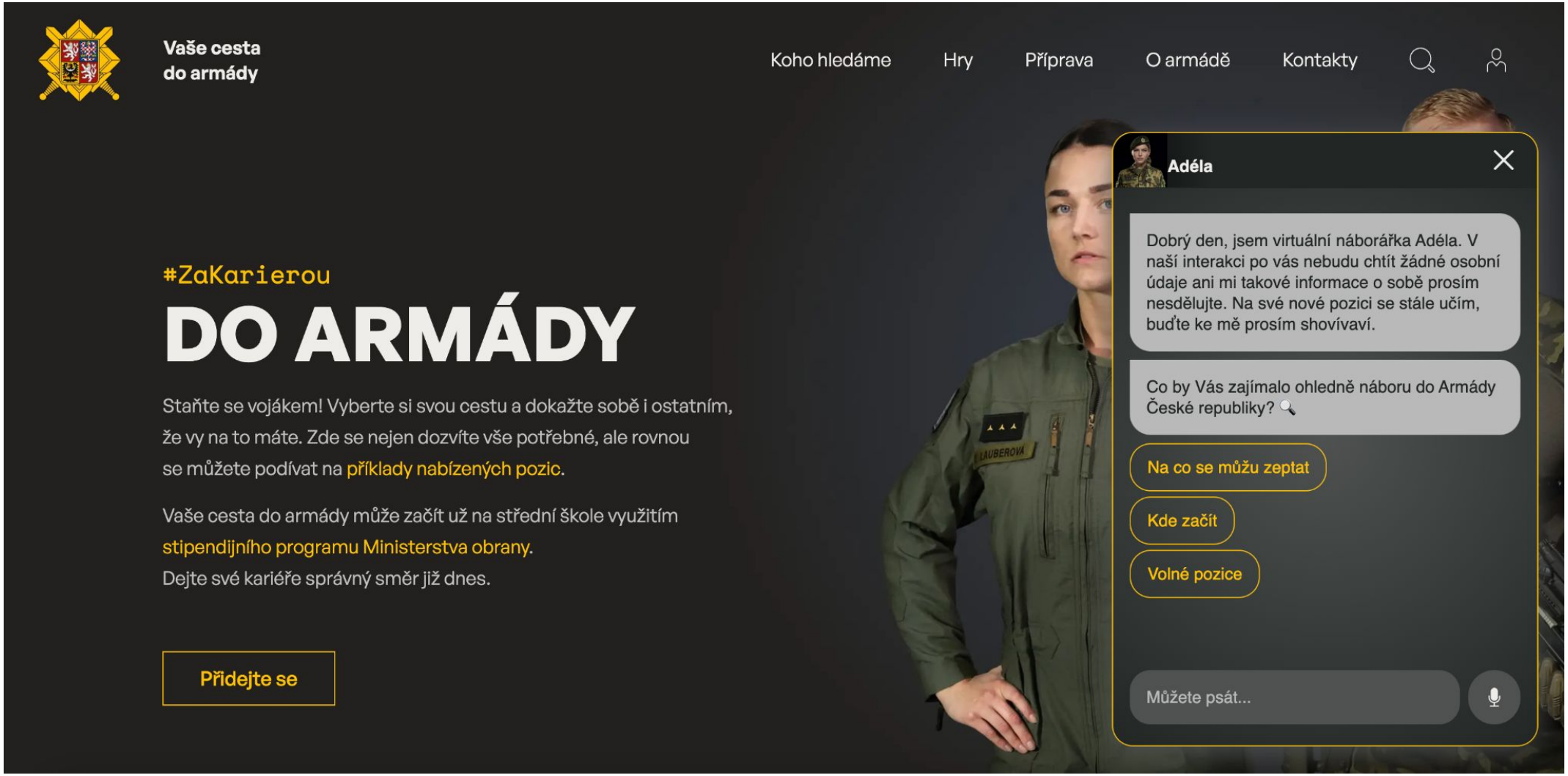
- Automatically
  - Learning from user selections
  - Statistics on user selections – automated "pre-selection" for next users
- Semi-automatically or manually
  - Chat log analysis → model update
  - Used Measures:
    - **Coverage** ... rate at which your chatbot is confident that it can address the user's request (per dialogue turn)
    - **Containment** ... rate at which your chatbot can satisfy a user's request without human intervention, i.e. connect to human agent not requested (per conversation)
- Metrics
  - Top 22 Metrics for Chatbot Analytics in 2024:

    https://research.aimultiple.com/chatbot-analytics/
- Testing was and is always a challenge (Botium)

Authoring a Dialogue System “new style”



Authoring a Dialogue System “new style”

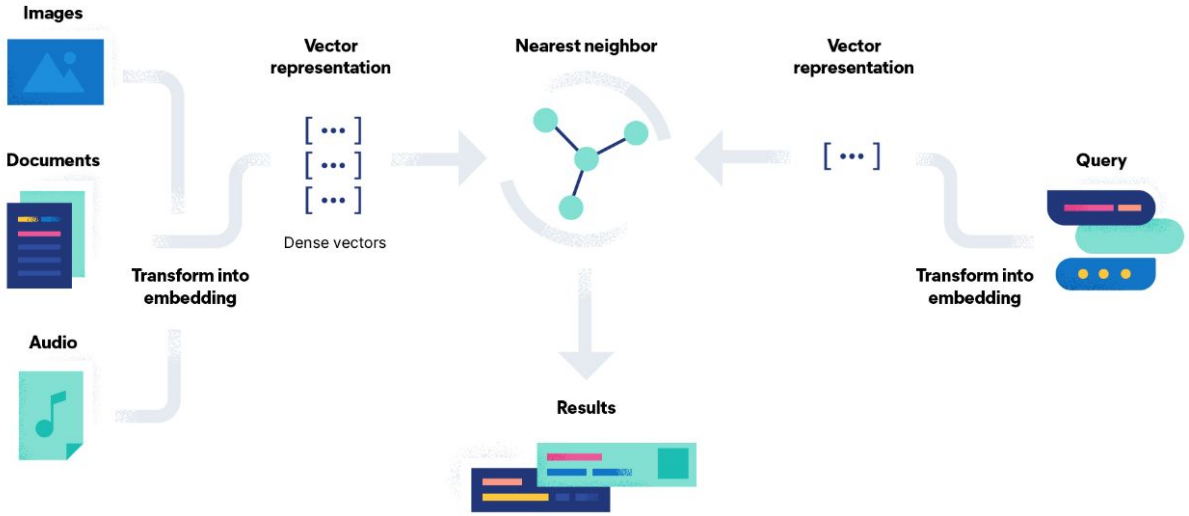


Semantic Search, RAG

- Semantic Search (mSearch)

  Semantic search

  ~ search technology that interprets the meaning of words and phrases



Source: Elastic https://www.elastic.co/what-is/semantic-search

- RAG (mDataChat)

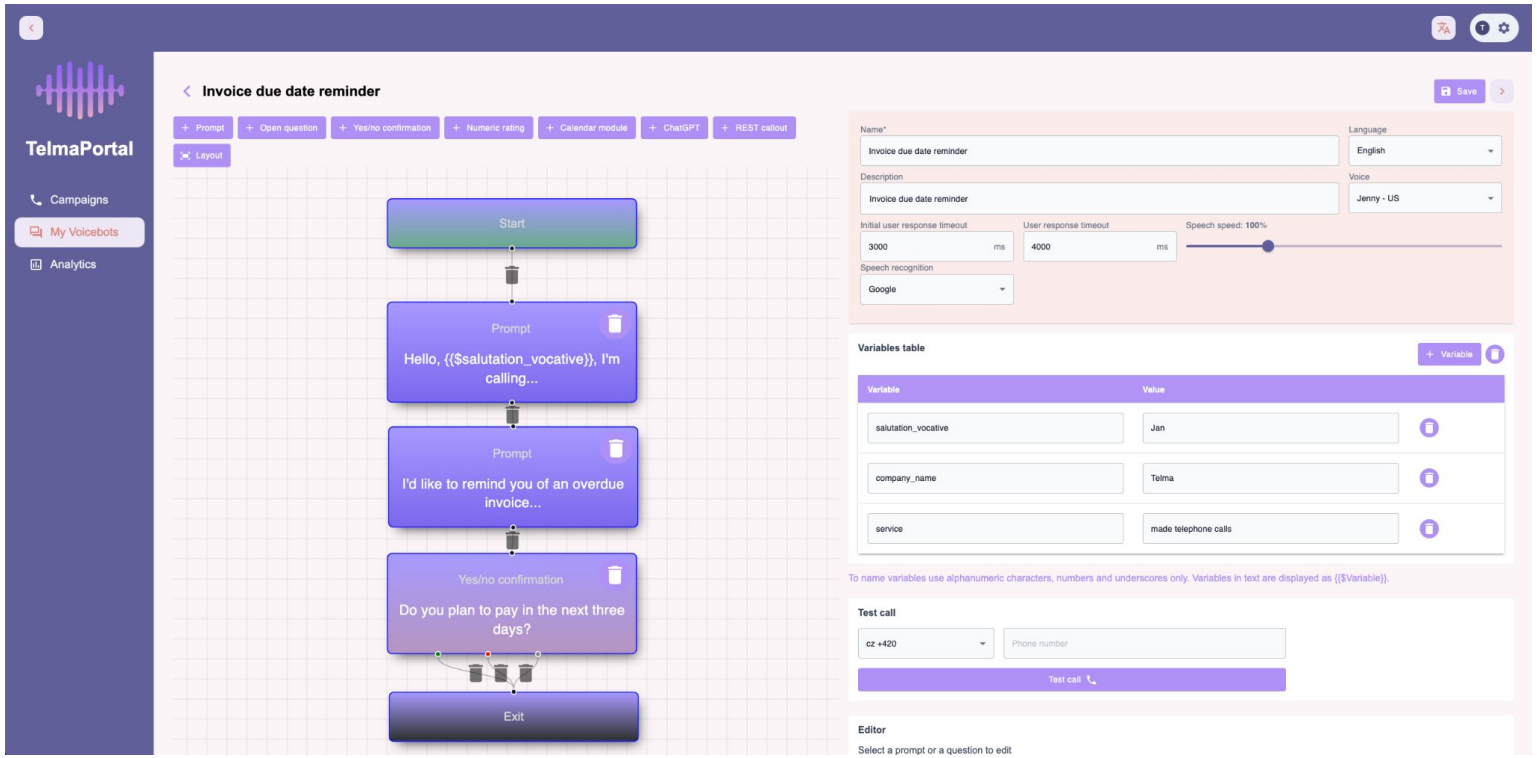
  Retrieval-Augmented Generation

  ~ process of optimizing the output of a large language model, so it references an authoritative knowledge base outside of its training data sources before generating a response

Authoring tools by Mama AI/Telma AI

- Mama Telma AI tooling for outbound calls
  - No-code/low-code
  - Modularization
    - Yes/No/Don’t know
    - Open question
    - Calendar
  - Webhook, SMS integration
  - GPT module, semantic search

Examples: https://telma.ai/products/outbound



- Voice lounge concept
  - Python implementation
  - Modules
    - (longer) Number dictation
    - Address dictation (RÚIAN)
  - Guess animal game on Alexa (see youtube)

Examples: https://telma.ai/products/inbound

Generative AI vs. Agentic AI

Key Features of Generative AI

Content Creation - Excels at generating essays, answers, code, and creative text. Tools like ChatGPT streamline writing and software development.

Data Analysis - Processes large datasets to find trends and patterns; enhances workflows and improves customer experience.

Adaptability - Adjusts output based on user input and feedback, refining results to better match expectations.

Personalization - Delivers tailored experiences and recommendations, especially valuable in retail and customer engagement.

Key Features of Agentic AI

Decision-Making - Assesses situations and acts on pre-defined goals with minimal human involvement.

Problem-Solving - Perceive–reason–act–learn cycle; integrates LLMs with tools for continuous learning.

Autonomy - Operates independently, enabling automation of complex tasks in organizational workflows.

Interactivity - Reacts in real-time to its environment (e.g., self-driving cars analyzing surroundings).

Planning - Executes multi-step strategies to reach specific objectives in dynamic scenarios.

Source: https://www.ibm.com/think/topics/agentic-ai-vs-generative-ai

Practical Examples

Thank you for you attention

Jan Cuřín <jan.curin@themama.ai>

https://themama.ai

https://telma.ai

Jan Cuřín