Explainable Voice & Conversational AI: Making Voicebots and Chatbots great again!

Dr Maria Aretoulaki

Everybody is talking nowadays about chatbots and Conversational AI, especially with the latest ChatGPT3 hype! The imagination of experts and the public alike has been captured by its human-like language, the eerie variability in its responses, its ability to understand the nuances of human language and even logic and creativity. Nevertheless, few people know or realise that chatbots and other conversational systems have been around for several decades!

The first chatbot was arguably ELIZA, the Rogerian Psychotherapist, in the 60s. It would talk about emotions, relationships, but it was only spotting keywords, there was no understanding behind it. You can try it out yourself.

ELIZA then and now

Then, in the 90s, the focus moved to spoken dialogue systems, the precursors of Speech IVRs (Interactive Voice Response systems) and what we would nowadays call “voicebots”. Still in Academia, very sophisticated voice systems were developed that would give train and air travel information or would help you navigate a map or negotiate a task. They were based on a combination of statistics and semantics: the statistical approach introduced robustness compared to the keyword-based approach, and the semantics was introducing some kind of knowledge representation (ontology) and a degree of discourse structure, both of which were contributing to actual “understanding”.

Around the year 2000, the IVRs that we all know and hate appeared mainly in Call Centre automation. You would get them when you called your Bank, Insurance or even Tax Office. Back then, these systems were quite crude as they were based on hand-crafted recognition grammars and equally manual, directed dialogues that were neither robust nor user-centred or flexible. A manual grammar would literally spell out how people say or ask for things along

with all predictable permutations; e.g. to get their account balance, a user was expected to either say "*I want my account balance*" or "*I need my account balance*" or "*My account balance please*" or "*erm account balance*" etc. If an actual user didn't use any of the prespecified linguistic expressions, the IVR would fail to recognise them and would awkwardly ask the user to repeat themselves, occasionally ad nauseam!

Manual Grammar for "No"

```
public <nophrases> = (
                      [actually] no | nope | nah |
                      [please] dont |
                      not (necessary | exactly) |
                      no need |
                      i dont need that |
                      [thats] (wrong | not (correct | right | the one [i (wanted | asked for)] ) )

                     [ <0_GeneralGrammars_Swearwords.swearwords> |
                       <0_GeneralGrammars_Politeness.endPolite>
                     ]
                    ) * ;
```

Manual Grammar for "Yes"

```
public <yesphrases> =
                       ( yes | yeah | yep |
                         okay | okey dokey |
                         correct | right |
                         thats (correct | right | the one | fine | okay) |
                         exactly | precisely |
                         <0_GeneralGrammars_Politeness.endPolite>
                       ) * ;
```

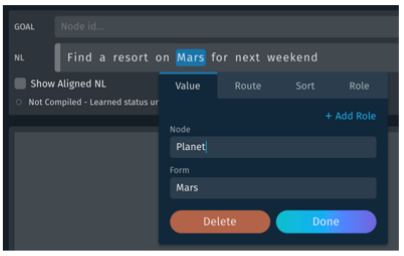
Around 2010, the first voice assistants appeared, first APPLE's SIRI and later the GOOGLE Assistant and the AMAZON Alexa and we all got excited about chatting with them as if they were our friends about Life, the Universe and Everything. Voice Assistants were using statistical grammars, so they were quite robust if you talked about something they already knew about, but would still fail miserably if you picked the wrong topic. By that time, the original crude IVRs had already been in use for a decade, which meant that a decade's worth of real-world speech data had also been collected to train and finetune speech recognition on. This profited both types of speech systems but more so the Call Centre IVRs because they were domain-specific and hence more restricted in the range of topics they could handle, which in a way was also guiding the users to cooperate, and hopefully get things done too.

Amazon Alexa

By 2020, Deep Learning Machine Learning (ML) approaches had been developed that could both mimic the complexity and multi-layered nature of learning in the human brain and also process the now ubiquitous terabytes of training data available. Deep Learning-based

voicebots and chatbots appeared that were also really robust, once again if you chose the right topic. Once again, domain-specificity brings the advantage that an out-of-the-box solution does not.

utterance transcription	action	topic	label	intent
my balance	enquire	balance	enquire-balance	enquire-account_balance
total balance	enquire	balance	enquire-balance	enquire-account_balance
what is my total balance	enquire	balance	enquire-balance	enquire-account_balance
ah can you please tell how much I owe you	enquire	balance_outstanding	enquire-balance_outstanding	enquire-account_balance
how much do I owe	enquire	balance_outstanding	enquire-balance_outstanding	enquire-account_balance
how much is my past due	enquire	balance_outstanding	enquire-balance_outstanding	enquire-account_balance
i'm calling about my bill	enquire	bill	enquire-bill	enquire-bill
how much is my bill	enquire	bill_amount	enquire-bill_amount	enquire-bill
e-bill	enquire	bill_paperless	enquire-bill_paperless	enquire-bill_paperless
paperless	enquire	bill_paperless	enquire-bill_paperless	enquire-bill_paperless
paperless billing	enquire	bill_paperless	enquire-bill_paperless	enquire-bill_paperless
direct debit	enquire	direct_debit	enquire-direct_debit	enquire-auto_pay
how can i make a payment	enquire	payment_options	enquire-payment_options	pay-bill_options
how do i make a payment	enquire	payment_options	enquire-payment_options	pay-bill_options
what are my payment options	enquire	payment_options	enquire-payment_options	pay-bill_options
make a payment	make	payment	make-payment	pay-bill
make my payment now	make	payment	make-payment	pay-bill
bill pay	pay	bill	pay-bill	pay-bill
pay bill	pay	bill	pay-bill	pay-bill
pay my bill	pay	bill	pay-bill	pay-bill
to pay my bill	pay	bill	pay-bill	pay-bill
can i pay my bill online	pay	bill_online	pay-bill_online	pay-bill_online
make a payment online	pay	bill_online	pay-bill_online	pay-bill_online
how do i pay my bill	pay	bill_options	pay-bill_options	pay-bill_options



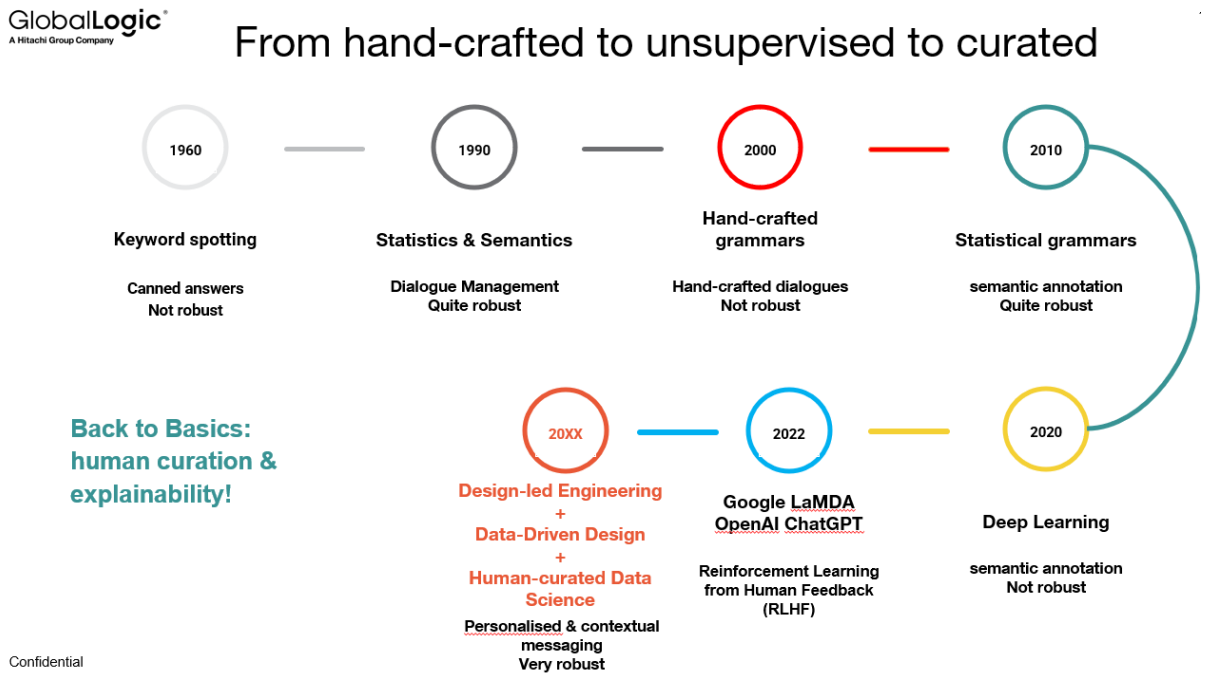
This evolution of voicebots and chatbots from the ‘60s to 2020 can be summarised as starting from Academic research to Commercial applications and culminating in today's ubiquitous commoditised general purpose smartspeaker and domain-specific Voice and Digital Assistants, and even the Digital Humans of the impending Metaverse apocalypse! This is where we are now 60 years later; the black box has become a white box and all language technologies (Speech Recognition, Natural Language Processing [NLP] and Natural Language Understanding [NLU]), as well as the AI and ML approaches behind them have been democratised.

This democratisation of the technology and the tools to build a natural language bot has meant that the human expert has been sidelined or outright removed. This is why nowadays most of the voicebots and chatbots we interact with on a daily basis are no longer designed by people whose expertise lies precisely in the intersection of Linguistics and Engineering, i.e. Voice User Interface (VUI) Designers, Conversational User Interface (CUI) Designers, Conversational Experience Designers (CxD) or at least Computational Linguists. Instead, the vast majority of Alexa Skills, Google Actions, Samsung Capsules and Customer Service bots in the market today have been produced by developers, who don't have the linguistic expertise to leverage how and why human language works, even if they are native speakers, and by copywriters, who don't have the technical knowledge or experience to predict when and why the underlying Language Technology fails.

Human Language is complex, unpredictable, ambiguous, imperfect and messy, and humans can process and use it as a communication tool, because they have intuitive knowledge of how multiple language components fit together and their relationship to the world around them: Phonetics and Phonology (sound), Morphology and Syntax (form), Semantics (meaning) and Pragmatics (intent and context). At the same time, Language Technologies are "inexact" Sciences, because Human Language itself is messy. Thus, Speech Recognition, NLP and NLU all largely depend on how good the training data is and whether, who and how curated this data. Language processing is guesswork really and Language understanding can only be done in context, i.e. taking into account the user's profile, history, physical environment, goals, expectations and assumptions.

Then of course in 2022 GOOGLE LaMDA emerged, with one Googler claiming that the chatbot had become “sentient”, which made Google quickly hide it away. Then Open AI won that race by making their ChatGPT3 bot publicly available to everybody’s amazement. Both are based on Large Language Models (LLMs), which are very robust for the processing of

real language, given the amount of real-world language training data they are based on. However, human feedback is part of how they learn; a human is expected to correct wrong answers or downgrade suboptimal answers. And it is wrong a lot. It has been proven to hallucinate facts and be very assured of its truthfulness, which is very dangerous. Suddenly the human is brought back in.



That's why we need to **go back to basics**: away from the extremes of frail hand-crafted but also completely unsupervised approaches, and move towards a semi-supervised human-curated approach that leads to the golden duo of robustness and explainability. **'Explainable Voice & Conversational AI'**, along the lines of Explainable AI, **combines** data-driven **Voice and Conversational UI and UX Design with Human expert-curated Data Science**. It ensures that we don't just do NLP but also NLU, we move from pure processing to actual understanding. The human expert needs to be trained in Computational Linguistics, Language Engineering, NLP and Speech Recognition and have experience working with and designing for massive amounts of real-world speech and text language data. **Language and Speech Analytics is a specialist type of Data Science and VUI, CUI and Conversation Design is a specialist type of UI and UX Design.** This hybrid approach ensures optimal language coverage and at the same time the transparency, standardisation, customisation and user-centred control of a rule-based system.

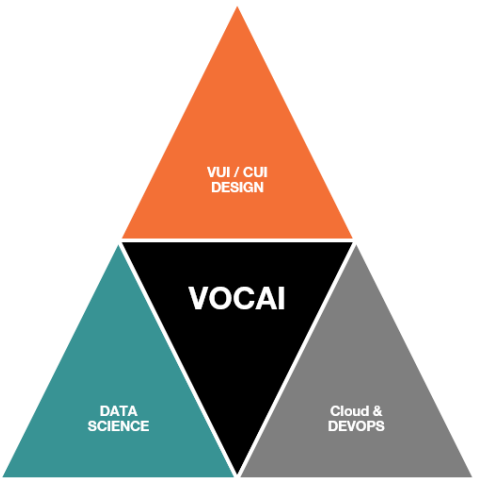
Design-led AND Data-driven AND Cloud-enabled

From NLP & ASR to NLU

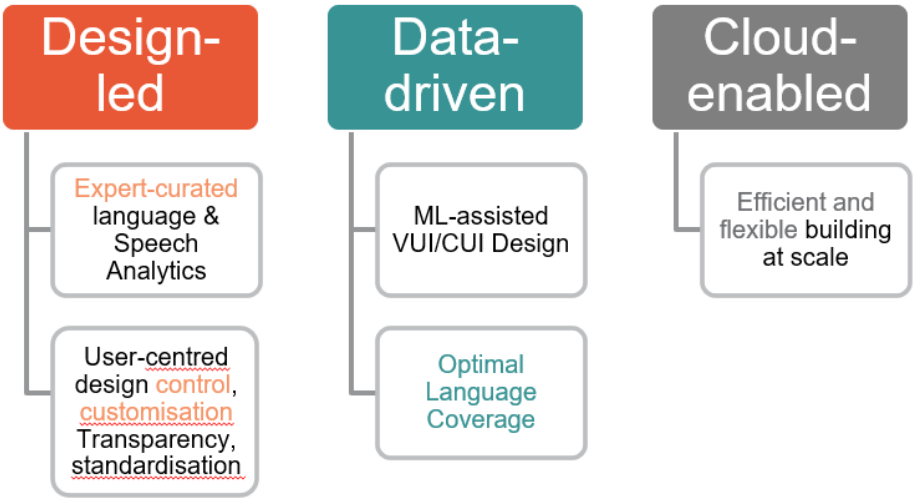
NLP: Natural Language Processing (word strings)

NLU: Natural Language Understanding (meaning & intent)

NLU: understanding what humans say, type & mean and how to better act on that in a **personalised, context-dependent and seamless** way (customers, employees, others)



This **3-pillar Methodology to Voice & Conversational AI (VOCAI)** leverages **Conversational UX Design, Data Science and Cloud & DevOps.** The approach guarantees the design and delivery of not just off-the-shelf commoditised NLP, but **client-tailored Natural Language Understanding (NLU)** for the delivery of personalised, context-dependent and seamless solutions and services.



A bit about the author

Dr Maria Aretoulaki is a Voice & Conversational User Interface & Experience Design veteran and pioneer. She has been designing natural language UIs, Spoken Dialogue Information Systems, Speech IVRs, Voicebots and Chatbots for the past 30 years. She has designed and optimised Conversational UIs for Contact Centre Self-Service applications working with multinational Tech giants, Telecoms, Retail and Investment Banks, Insurance, Security and Logistics companies, Utilities and Governments, as well as voice platform providers, CRM and Contact Centre technology providers and Systems Integrators.

In 2018 she coined the terms 'Explainable VUI & CUI Design' and 'Explainable Voice & Conversational AI' to advocate for Natural Language Dialogue Interface Design and Development that combine Computational Linguistics and NLU expertise with real-world Language Engineering and Speech Analytics experience. She joined GlobalLogic UK&I in 2022 as a Principal Consultant Consumer Solutions & Experience, where she heads our Voice & Conversational AI CoE.