Introduction to Conversational AI

Where Computational Linguistics (CL) meets Human-Computer Interaction (HCI)
Who has had a conversation with a machine?
Three examples:

1. Telephone hotline bots
2. Smart phone voice assistants
3. Social robots
Lecture take-away:

- Speech is more than spoken text
- Only few aspects of talking can be automated
- Everyone, whatever your background, can contribute to the field of Conversational AI
What is Conversational AI?

>>the task of building machine that can verbally interact with humans<<

- Accelerated through advances in Automatic Speech Recognition (ASR) in the 2010s
- Launch of Apple’s “Siri” voice assistant in 2011
- Google duplex demo in 2018
- Recent social robot hype: *Sophia, Furhat, Pepper, Nao* ..
A new way to interact with computers

From...
Command line interface
to...
Graphical user interface (GUI)
to...
Voice user interface (VUI)
The voice interface: the user perspective

- can be more intuitive, enabling people to use the most natural means of communication
- enable users to complete certain tasks faster than by using a GUI
The voice interface: the **industry** perspective

- *improved user experience*
- *enable new products*
- *more personalized interaction with technology*
- *reduce costs*
The voice interface: the research perspective

- **scientific challenges**
  basic and applied research on social interaction and communication need

- **design challenges**
  new design language needed to implement this new type of interface

- **technical challenges**
  how to build new tools for more robust/advanced speech processing

- **ethical and privacy issues**
  new ethical and legal frameworks needed to regulate this powerful new and potentially invasive technology
Where **CL meets HCI meets Conversation Analysis (CA)**

➢ Challenge of the CL/HCI meeting point: the discrepancy between what is known about talk and what is technically possible

➢ Challenge of the HCI/CA meeting point: how can we design and build voice technology informed by Conversation Analysis - the science of talk.
How does voice technology work?

A speech processing pipeline: ASR, NLU and TTS modules

Automatic speech recognition (ASR), end-to-end

From *hello* the word... to “hello” the string

Automatic speech recognition (ASR), end-to-end

“Printing” sound: the Fourier / Wavelet transform

From soundwaves to spectrograms - from audio signal to graphical representation

**Automatic speech recognition (ASR)**

The Tchaikovsky problem:
Saying “Tchaikovsky” often ends up being transcribed as “chai coffee” or “shall I cough sky”

**Current state of the art**

Text-to-speech (TTS)

“Reverse ASR”: from strings to soundwave

Some current issues: natural intonation, latency, no reflexive online attention to user behaviour

Natural language understanding (NLU)

ASR input ---> process “meaning” ---> TTS output

Natural language understanding (NLU)

Example:

“I want a pizza
with ham and
extra pineapple”

Natural language understanding (NLU)

Example:

“I want a pizza with ham and extra pineapple”

TakeOrder{
  Type= pizza,
  Topping: [ham],
  Extra: [pineapple]
}

Current state of the art

Natural language understanding (NLU)

Example:

“I want a pizza with ham and extra pineapple”

```
TakeOrder{
  Type= pizza,
  Topping: [ham],
  Extra: [pineapple]
}
```

if TakeOrder=True:
say response[RepeatOrder]
else:
start_seq[TakeOrderFallback]

“Got it, your order is one pizza with ham and extra pineapple.”

Natural language understanding (NLU)

Example: ...but what if the user instead says:

“I want a pizza with ham and extra pineapple”

“One pizza with that tasty ham and also pineapple, please”
Natural language understanding (NLU)

Rule based: *define a set of rules to determine pizza type*

Keyword spotting: “I want a pizza with ham and extra pineapple”

“One pizza with that tasty ham and also pineapple, please”

*NLU processor:* [intent] [type] [topping] [topping type 1] [topping type 2]
Natural language understanding (NLU)

Machine learning based: compute the rules based on data

Training dataset:

1 “I want a pizza with ham and extra pineapple”
2 “I want pizza and loads of gammon and pineapples”
3 “Give me one with ham and also pineapples”
4 “Pineapple ham pizza please”
5 “Ehh pineapple pen pizza please”
6 “I.. uhm.. I want a pizza with... ham and olives, no pineapples”

Natural language understanding (NLU)

Machine learning based: *compute the rules based on data*

Training dataset:

1 “I want a pizza with ham and extra pineapple”
2 “I want pizza and loads of gammon and pineapples”
3 “Give me one with ham and also pineapples”
4 “Pineapple ham pizza please”
5 “Ehh pineapple pen pizza please”
6 “I.. uhm.. I want a pizza with... ham and olives, no pineapples”

User says:

“One pizza with ham and pineapple, thanks”
Natural language understanding (NLU)

Machine learning based: *compute the rules based on data*

Training dataset:

1. “I want a pizza with *ham* and extra *pineapple*”
2. “I want pizza and loads of gammon and *pineapples*”
3. “Give me one with *ham* and also *pineapples*”
4. “*Pineapple ham* pizza please”
5. “Ehh *pineapple pen pizza please*”
6. “I.. uhm.. I want a pizza with… *ham* and *olives*, no *pineapples*”

<table>
<thead>
<tr>
<th>Pizza-Order-Type classifier:</th>
<th>Probability</th>
<th>Pizza types:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Probability</strong></td>
<td><strong>Pizza types:</strong></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>Ham+Pineapple</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>Anchovies</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>Seafood</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>Ham+Olives</td>
<td></td>
</tr>
</tbody>
</table>

User says:

“One pizza with *ham* and *pineapple*, thanks”
Natural language understanding (NLU)

...but what if the user instead says:

“One pizza Hawaii please” 😲
Natural language understanding (NLU)

...whether you define the rules for sorting or compute them using machine learning, in practice, NLU is..

- a sorting problem
- utterance to (intent) label mapping
- a state-transition computation
Natural language understanding (NLU)

Chatbots as state machines:

Example architecture of a chatbot:

Credit: https://towardsdatascience.com/architecture-overview-of-a-conversational-ai-chat-bot-4ef3dfe6d52e?gi=165126e5cf47
Limits of Natural language understanding (NLU)

- A tool to build useful tools rather than understanding language in any human sense
Limits of Natural language understanding (NLU)

- A tool to build useful tools rather than understanding language in any human sense

Image building a concierge bot...
Limits of Natural language understanding (NLU)

➢ A tool to build useful tools rather than understanding language in any human sense

Image building a concierge bot... then someone comes and puts a hat on it!
Limits of Natural language understanding (NLU)

➢ A tool to build useful tools rather than understanding language in any human sense

Image building a concierge bot... then someone comes and puts a hat on it!

➢ NLU tools are limited to specific use cases, don’t “scale up” well, imagine...
  ○ a generic model for politeness/rudeness?
  ○ a generic model of ‘common sense’?
  ○ a generic model of building rapport?
Limits of Natural language understanding (NLU)

➢ A tool to build useful tools rather than *understanding* language in any human sense

Image building a concierge bot... then someone comes and puts a hat on it!

➢ NLU tools are limited to specific use cases, don’t “scale up” well, imagine...
  ○ a generic model for politeness/rudeness?
  ○ a generic model of ‘common sense’?
  ○ a generic model of building rapport?

➢ Problem of situated interaction
Limits of Natural language understanding (NLU)

➢ Problem of situated interaction: (see also 'symbol grounding problem')

Example: Image building a Pizza bot here...

...and in Italy
Limits of Natural language understanding (NLU)

➢ Problem of situated interaction: (see also ‘symbol grounding problem’)

Example: Image building a Pizza bot here...

...and in Italy

Requires grounding the concept of “pizza” in local context; Bots don’t generalize well, need localization to environment
Some conceptual issues:
- Searle’s Chinese room argument
- the antrophomorphization of technology
Some conceptual issues:
- Searle’s Chinese room argument
- the anthropomorphization of technology

There are plenty of engineering contests on mimicking human verbal behaviour: Turing test, Winograd schema challenge, Loebner prize, Alexa prize etc.
Some conceptual issues:
- Searle’s Chinese room argument
- the anthropomorphization of technology

There are plenty of engineering contests on mimicking human verbal behaviour: Turing test, Winograd schema challenge, Loebner prize, Alexa prize etc.

Can we even “compute” talk?
Some conceptual issues:
- Searle’s Chinese room argument
- the anthropomorphization of technology

There are plenty of engineering contests on mimicking human verbal behaviour: Turing test, Winograd schema challenge, Loebner prize, Alexa prize etc.

Can we even “compute” talk?

Bottom line: No machine can currently participate in human social life in a convincingly human-like way :( 

Some conceptual issues:
- Searle’s Chinese room argument
- the anthropomorphization of technology

There are plenty of engineering contests on mimicking human verbal behaviour: Turing test, Winograd schema challenge, Loebner prize, Alexa prize etc.

Can we even “compute” talk?

Bottom line: No machine can currently participate in human social life in a convincingly human-like way :( ... or wait... maybe it’s better this way

Remember Eliza?

It is easy to build interactive software that “deceives” humans into believing it possesses human qualities or intelligence.
Remember Eliza?

It is easy to build interactive software that “deceives” humans into believing it possesses human qualities or intelligence..

Weizenbaum himself on ELIZA: a PARODY of a therapist, a fake, a charade, aim was to illustrate the LIMITS of computers, not their potential (Weizenbaum 1966)

Two fundamental problems:

*How to formally represent units in talk?*

*How to formally represent “meaning” in talk?*
Current state of the art

Formal representations of talk?
How to represent talk in computer-readable formats:
Formal representations of talk?

How to represent talk in computer-readable formats:

Movie subtitles

Current state of the art
Formal representations of talk?

How to represent talk in computer-readable formats:

- Movie subtitles
  - दूर हो जाओ, मुझसे दूर हो जाओ
  - Get away! Get away from me!

- Phonetic transcripts
  - (IPA, ARPAbet)
  - I've been staring at the edge of the water
  - sir bi'n ster-jing at d3 eijv d3 w3t3r
Formal representations of talk?

How to represent talk in computer-readable formats:

- Movie subtitles
- Phonetic transcripts (IPA, ARPAbet)
- Transcripts+Mark-up language (CHAT, XML, SSML)

Current state of the art

"{\'input\': {\'ssml\': \"<speak>The <say-as interpret-as="characters">Welcome</say-as> standard <break time="1s"/> to the <sub alias="matrix">emph=1</sub> .</speak>\",\
'voice':{"languageCode":"en-us",
'name':'en-US-Standard-B',
'ssmlGender':'MALE'},'audioConfig':{'audioEncoding':'MP3'}}"
Formal representations of talk?

**Challenges:** How to systematically represent...
Formal representations of talk?

**Challenges:** How to systematically represent...
Non-lexical vocalizations?
Formal representations of talk?

**Challenges:** How to systematically represent...

- Non-lexical vocalizations?
- Phonetic reductions?
Formal representations of talk?

**Challenges:** How to systematically represent...

Non-lexical vocalizations?

Phonetic reductions?

What is noise and what is signal? Stretches, stress, pitch, intensity? Which phonetic properties need to be included?
Formal representations of talk?

**Challenges:** How to systematically represent...

Non-lexical vocalizations?

Phonetic reductions?

What is noise and what is signal? Stretches, stress, pitch, intensity? Which phonetic properties need to be included?

Many subtle cues in social interaction are often not processed at all.
Reducing talk to strings and tags:

Limits of representing talk are most visible on the micro and macro ends of the unit spectrum in linguistics. Consider...
Reducing talk to strings and tags:

Limits of representing talk are most visible on the micro and macro ends of the unit spectrum in linguistics. Consider...

**Phonetic/phonological units in talk:**
“uhmm”, “err”, haha, <lip smack> (e.g. Keevallik and Ogden 2020)
Reducing talk to strings and tags:

Limits of representing talk are most visible on the micro and macro ends of the unit spectrum in linguistics. Consider...

**Phonetic/phonological units in talk:**
“uhmm”, “err”, haha, <lip smack> (e.g. Keevallik and Ogden 2020)

**Pragmatic units in talk:**
Reducing talk to strings and tags:
Limits of representing talk are most visible on the micro and macro ends of the unit spectrum in linguistics. Consider...

**Phonetic/phonological units in talk:** “uhmm”, “err”, haha, <lip smack> (e.g. Keevallik and Ogden 2020)

**Pragmatic units in talk:**
???
Reducing talk to strings and tags:

Limits of representing talk are most visible on the micro and macro ends of the unit spectrum in linguistics. Consider...

**Phonetic/phonological units in talk:**
“uhmm”, “err”, haha, <lip smack> (e.g. Keevallik and Ogden 2020)

**Pragmatic units in talk:**
Proposed units: turns, turn-constructional unit (TCU), maybe clause
(The idea of ‘speech acts’ or what is an ‘intent’ is still hugely controversial in the literature, see e.g. Button 1994)

Formal representations of “meaning”: some examples
Formal representations of “meaning”: some examples

90s: Formal cognitive/“deep semantic” models of language

Lexical databases of formal representations of cognitive schemas, e.g. FrameNet (Baker et al. 1998)
Formal representations of “meaning”: some examples

90s: Formal cognitive/“deep semantic” models of language

Lexical databases of formal representations of cognitive schemas, e.g. FrameNet (Baker er al. 1998)

00s: Formal speech act taxonomies

Formal models of what people are doing in talk, such as accusing, asserting, apologizing, building arguments and narratives (cf. DAMSL, Dialog Act ISO (Core and Allen 1997, Bunt et al. 2010))
Formal representations of “meaning”: some examples

90s: Formal cognitive/“deep semantic” models of language

Lexical databases of formal representations of cognitive schemas, e.g. FrameNet (Baker et al. 1998)

00s: Formal speech act taxonomies

Formal models of what people are doing in talk, such as accusing, asserting, apologizing, building arguments and narratives (cf. DAMSL, Dialog Act ISO (Core and Allen 1997, Bunt et al. 2010))

10s: Statistical Language Models

Neural language models for dialog (“stochastic parrots”): DialoGPT, ConveRT (Zhang 2020, Henderson et al. 2020, Bender et al. 2021, Bender and Koller 2020)
Formal representations of “meaning”: some examples

90s: Formal cognitive/“deep semantic” models of language

Lexical databases of formal representations of cognitive schemas, e.g. FrameNet (Baker et al. 1998)

00s: Formal speech act taxonomies

Formal models of what people are doing in talk, such as accusing, asserting, apologizing, building arguments and narratives (cf. DAMSL, Dialog Act ISO (Core and Allen 1997, Bunt et al. 2010))

10s: Statistical Language Models


No formal framework can represent meaning in natural conversation. The flexibility of language-in-interaction and in-situ meaning-making practices continues to pose a critical challenge in computational linguistics and NLP.


The limits of automation: from NLP to HCI

What cannot be formalized, cannot be automated...

...and has to be built by-hand.
The limits of automation: from NLP to HCI

What cannot be formalized, cannot be automated...

...and has to be built by-hand.

This is where conversation design / interaction design comes in.
Conversation design (CD):

CD is a design discipline derived from UI/UX design and copywriting

- towards better (more useful / more human-like) conversations with bots, make human-bot interaction more successful

CD is a design language based on human conversation (similar to how material design is a design language based on pen and paper)
Conversation design (CD):

CD is a design discipline derived from UI/UX design and copywriting

- towards better (more useful / more human-like) conversations with bots, make human-bot interaction more successful

CD is a design language based on human conversation (similar to how material design is a design language based on pen and paper)

CD is the practical solution for the many “scaling problems” that beset automation in Conversational AI
Human-computer interaction

Build dialogs

Which voice?

Which bot persona:
Conversation design (CD) for voice bot development

Design for voice user interfaces (VUI) in voice bots:

Range of NLU software: Google dialogflow, Amazon Alexa skills, wit.ai, Rasa...

Part of chatbot development teams

Bot's life cycle

- Design
- Build
- Test
- Publish
- Connect
- Evaluate
Summary: Talking machines?
Summary: Talking machines?

Not happening.. but

While technical, conceptual and practical challenges remain...

...current voice interfaces are a significant technological achievements and enable useful products
Summary: The Conversational AI industry

**Conversational AI engineer:** develop and implement tools related to speech processing / NLU pipelines, push the limits of what’s technically possible

**Conversation designer:** design and prototype human-bot interaction, improve user experience, examine dialog flows and error sources, voice branding, persona design...
Ongoing research at PolyU

★ Empirical and experimental studies: representations of conversations, coordination and prediction for turn-taking (using Furhat, CoBra network, forthcoming)
★ Technical studies: moving away from intent-based NLU (cf. Sankar et al 2019)
★ Bot skill development for Chinese Conversational AI (Liesenfeld and Huang 2020)
★ Design studies: Conversation design for Chinese (collab. with School of Design)

If working with Social robots and Conversational AI has peaked your interest, there are opportunities for students to join such current projects at PolyU :)

Now lets get building!
Build an interactive bot:

Use “Rasa playground” to build a useful bot

Put your knowledge of NLU to use:

- define training data, intents, entities, responses and dialog states

Play the role of both a conversation engineer and designer:

- Design, build and test your bot
- Play around with different conversational styles and different dialog flows
Details: build a bot that automates pizza orders

➔ the bot should understand requests for 3 different types of pizza
➔ The bot should have at least three states:
  ◆ “return greeting”
  ◆ “take and confirm pizza order”
  ◆ “say goodbye”
➔ Define training data, intents and entities for each state
➔ Train and test your bot
Today’s lecture aimed to provide you with:

➔ a sense of design and implementation techniques for conversational agents
➔ an awareness of current research and emerging issues in the field of Conversational AI
➔ a introduction to the practical skills involved in building working conversational interfaces