UNIVERSITY OF LEEDS

Chatbots and Dialogue Systems

Professor Eric Atwell: Hello. This is Eric Atwell and this lecture I'm going to be talking about chatbots, also known as dialogue systems, and other things too. So, first of all, we're going to have an introduction to chatbots and dialogue systems and then have a look at properties of human conversation, since chatbots are supposed to mimic or emulate humans, and then look at a couple of ways of building these chatbots. One is rule based system.

And we'll look at some examples, ELIZA, PARRY, the AIML, artificial intelligence markup language, and then a couple more recent chatbots, ALICE and Hubert. And then another way to do this is corpus-based chatbots, using a corpus to train via some sort of machine learning, and then look at the more complex dialogue state architecture for some current chatbot platforms. And to end up with, look at a couple of issues evaluating dialogue systems or chatbots and ethical issues with chatbots.

OK, so let's start off by a general introduction. First of all, as with computational linguistics or text analytics or natural language processing, there are different names for this topic area. Conversational agents, also called dialogue systems or so-called dialogue agents. The general term seems to be chatbots. From different perspectives, an agent suggests that they are intelligently trying to do something or achieve some goal. Conversational sounds more like chat, whereas dialogue is more like interaction to some purpose.

Chatbots, bots means a robot of some sort. And you're probably familiar or have at least seen personal assistants on phones or other computing devices, like SIRI for Apple, Alexa from Amazon, Cortana from Microsoft, Google Assistant... I don't think they've got a fancy name for Google Assistant. You just say, hey, Google. All of these seem to be chatbots. In addition, you can...to talking to them via typing, you can actually literally talk to them, and they will give you a verbal response, an oral response.

However, we don't really know what's going on behind the scenes. So, these are apparently very smart systems, but on the other hand, if you try hard enough it's easy enough to ask questions which they just don't seem to be able to answer properly. They are aimed at a very general audience. They're not for specific purposes, but they have got a number of what are called skills or things like that, which are things they can do.

So, they're particularly good for playing music. I can say, Alexa, play me my favourite Sex Pistols song. Now try to guess which one it is. It can set timers. Alexa, tell me in an hour’s time. Clocks, various functions that they think people will need to do. It can also do chatting for fun. You can say, Alexa, tell me a joke.

There are some more, if you like, serious things you can do...booking travel reservations, ordering, shopping via Amazon if you're using Alexa. They can also answer questions, at least general knowledge questions. And typically, Alexa, if it hasn't got...if it's not a very straight forward answer, it will say, here's something I found in Google or Wikipedia or whatever. And clinicians, hospitals, clinical researchers have started to use these for mental health issues.

So, if you have mental health problems, one of the things that the patient has to do is talk to a clinician or a mental health practitioner, or a chatbot. And you can either just simply let it talk with the chatbot and hope that will improve things, or you can actually keep a transcript of what's being said and then afterwards analyse the transcript to do some sort of diagnosis. This does suggest there are two kinds of chatbots.

There's your chatting conversational agents which mimic informal human chatting, generally just for fun or for relaxation or for giving an individual a sense that they've got someone with them. And this potentially has therapeutic effects if you have mental health problems or are feeling lonely. But there are also the more formal targeted task-based dialogue agents, which have a particular task in mind like booking flights or booking restaurants or buying things via Amazon.

There are also interfaces to personal assistants. So, it would be very nice if every student, every academic had a personal assistant to deal with things like when is my next lecture or when is the

coursework due, and there are things to help with that sort of thing. Also, when driving a car, turn on the air conditioning or giving robot instructions or even given washing machines instructions.

Another sort of area for chatbots is question answering, ask a question. It's not clear if this counts as a conversational agent or a task-based agent, but there are certainly...for Amazon I can ask questions and sometimes it's more like human conversation. It's not particularly fact-finding questions. But if I'm trying to do some coursework, maybe I do need very task-oriented questions.

OK, so that's two types of chatbots. There are also, in general, two types of chatbot architectures and this also mimics the general sort of two types of architectures for AI systems and natural language processing systems in general. You can either have rule-based systems where the computational linguist writes some rules, tries it out, and extends the rules until eventually the system works reasonably well. Or you can have machine learning from a corpus where you have a training corpus appropriately annotated and the machine learning algorithm extracts, essentially, rules like the ones that the human did in the first version.

So, very simple rule-based systems have patterns and actions. So, there's a pattern and then the user input is matched against this set of patterns until the best pattern is found and then the corresponding action with that pattern is done. And the pattern can often be a question type and the action will be an answer, so that fits very well for question answering systems. In addition to that, you can stick on top of that some sort of infrastructure.

As we'll see in a minute, PARRY has a mental model, it’s responses aren't just based on the input patterns, but also its state of mind. And this actually produced transcripts which psychologists examined, and they weren't able to tell if PARRY was actually a human or a chatbot and, in some sense, it passed the Turing test. The Turing test is can human judges decide if they're conversing with an AI agent or a real person, and the psychologists weren't sure.

Corpus based system. Well, as we saw before, there's two sorts of general approaches. One is an information retrieval approach, encoding the input into some sort of vector and then finding in the

corpus the closest, the most similar vector using information retrieval techniques, and then outputting that. And the alternative nowadays is to use deep learning neural models to find the best match.

OK, so here's some examples into action. We've learned the bot a neural network system. "Will you sing me a song? Sure. What do you want me to sing to? I can sing a song about baking". And this looks like a plausible sort of interaction. Notice that the BlenderBot doesn't actually sing, so then the bot can't sing. It can simply, given the input, find an appropriate match in the training corpus and output that and it happens that in the training corpus there was probably some interactions where there were baking and singing going on. That's why it decided to volunteer to sing a song about baking.

Notice also that from the transcript, it's quite hard to see what the underlying architecture was, and this is certainly true for SIRI and Cortana and Alexa. You can try things out and these companies rely on the fact that people on the whole aren't that inquisitive about how the underlying algorithm works. They just want it to apparently work.

And here's this Chinese system developed by Microsoft China. I think it's called Xiaoice. Something like that. I'm afraid I can't pronounce this. Watch the video and you'll hear how it's pronounced. OK. And again, you can see this is a chatbot very widely used.

And there's actually a paper in The Computational Linguistics Journal, The Journal of the Association for Competition Linguistics describing in great detail more about the underlying algorithm, and it's for chatting to ordinary people in Chinese. When you start the app, you are advised you are not you're chatting to an AI agent, not a real person, but quite a few users go on for quite long conversations and think perhaps it is a person.

So, that's a couple of examples for task-based dialogue agents. They're not just simply chatting, so we don't necessarily have to have a huge training corpus covering all things that someone might say. But it is usually first focused on a task like setting a timer or making a travel reservation, so you have a model, potentially quite a small model, on a very specific topic.

And of course, what Alexa does, for example, is you have skills. Each of these skills is very much based on a goal or task, but you can basically power up as many skills as you want or have available several skills at the same time. So, that's how you do it. You basically have components or widgets for each of the skills and the more and more of these have, the more generic or general system has.

So, the architecture for a particular skill is to basically have a frame with slots and values, a sort of knowledge structure representing user intentions rather than just recording words and phrases. So, the frame may be a set of slots that the system tries to build in or fill, and the assumption is that in the interaction the user wants to fill the unknown slots and the system can help in filling the unknown slots. Each is associated with a question to the user.

So, for example, if you're trying to build a travel reservation, then the system knows that it has to have an origin, a destination, departure time, a departure date, and an airline if it's an airline booking system. So, in making the reservation, the system has to fill these slots and the user also knows probably that these are the bits of information you have to be given, so they're typical questions that the system can ask to try to fill each of these slots.

OK, so that's the general overview of what sorts of chatbots there are. There are conversational chatbots or task-based systems, and the architectures are either rule-based systems or corpus trained machine learning systems and they're either general conversations or they're focused on particular tasks. Now all of this, for a human user, they expect the conversation to be human-like.

So, this is a big problem that if you have a graphical user interface, then the graphical user interface designer can pretty much decide whatever they want to put on it, knowing that the human will realise that this is a graphical user interface. It's not a real person. But when you start conversing with a chatbot, you start making assumptions that it's going to talk in a natural human way. Here's a typical telephone conversation between a human travel agent A and a human client C.

And we'll see it's not just question, answer, question, answer, question, answer, but there are all sorts of strange extra things happening. OK, I'm going to point out a few of the examples. You don't

actually have to memorise this to pass the test at the end. OK. You see, in general there are these turns, but they're not simply question followed by answer followed by question followed by answer.

It is a sort of game where two people take turns. In fact, in a real conversation you may have more than two people in and they all take turns. A turn typically is a sentence, but it can just be a single word like OK, or it can be quite a longer sentence. Here's OK and you can respond OK just to say that it's my turn so I'm going to say something, but I'm just going to agree with you and that's it. Now you can take over your turn again.

Well, sometimes, this turn can be quite long. The client says, "what are they?" And the agent has to give quite a few possible options. And if it's turn taking, you have to then know when it's your turn...when do you take the floor? Or when is it the other person's turn, when do you yield the floor?

So, there's typically unspoken conversation. There's pauses or there is something in the intonation, the way the voice goes up or down which we learned to interpret in this way. You can also have interruptions. Notice agent 16 and client 17. The client starts talking before the agent is finished. He interrupts. And the human agent knows the interruption is coming, so he should stop talking. And he also knows that the client is about to make a correction.

So, the system has to be able to do this too. Now the agent says OK, there's two non-stops and at the same time...that hash means that they're happening simultaneously...at the same time, the agent's saying two non-stops. The client is saying, actually, and he's going to say, what day of the week is the 15th? So, he's going to change his date from the 15th possibly.

So, in human interactions, you are allowed to barge in. You want to allow the user to interrupt. And there's also this end-pointing task for a speech system of deciding whether the user has stopped talking. You can't really just wait for a long pause otherwise you have lots of long pauses in the interrupt. People do also pause in the middle of turns while they're thinking of what to say. So, you can sort of model the interaction as a series of turns, and each turn is a kind of action.

So, philosophers and linguists like Wittgenstein also have thought of it as being a dialogue act or a speech act, and there are types of these speech acts or dialogue acts as constatives committing the speaker to agreeing to something. OK? I concur with you. That's not actually introducing any new information other than I am agreeing with you. Or as a directive, trying to get someone to do something. Attempts by the speaker to get the addressee to do something.

I ask you; I'm asking you or forbidding you. There's also commissives, committing the speaker to do something themselves. I promise to do this. They can also be acknowledgments, expressing the speaker's attitude regarding the hearer. I am sorry. Hello. I thank you. So, these are different types of acts.

And if you go through a dialogue corpus of spoken dialogue, you will see there's plenty of examples and you can try annotating each of these things. This is a bit like part of speech tagging or named entity tagging. In this case, a whole piece of speech is labelled or tagged as being a directive or a constative or acknowledgment. And in a chatbot interaction, the same sort of thing should be true. You shouldn't simply have question followed by answer, question followed by answer, but you have to have these things in too like OK, which is an acknowledgment.

There's also the grounding principle. Participants in a conversation need to have a common ground. The principle of closure. Agents performing an action require evidence, sufficient for current purposes, that they have succeeded in performing it. So, if you're having a speech act, then each of the speakers need to be sure that they understand what's going on so far and they acknowledge that the hearer has understood.

That's why you have to say, OK, every now and again, or some sort of sound, like uh-huh to say, yes, I understand what you said so far and continue. You might think this is just in speech, but actually, this is a common thing in human computer interfaces. For example, you may have noticed in an elevator when you press the button it lights up. And you might think, why do that?

Well, it's to show, to acknowledge that the elevator does know that it has been called. If you press the button and nothing happens, then you're likely to keep pressing the button. But when light comes on, you know to stop pressing the button. And it's just that this is like an, OK, I understand action.

So, here we have an example. "You said, returning on May 15? Uh, yeah, at the end of the day. OK". You might have, "OK. I will take the 5ish flight on the night before on the 11th. On the 11th? OK". So, you can have just OK, or you can even repeat something...repeat the part of the utterance that you are confirming, that the 11th is the important bit in what he just said.

And "I need to travel in May. And, what day in May did you want to travel?" And means I acknowledge that you need to travel in May but in addition to that, I need to have some further specification. So, conversations have this structure. And they have, as I said, turn taking, or you might say adjacency pairs.

So, question and answer is the most common pair of things, but you can also have a proposal followed by an acceptance or rejection or a compliment followed by a downplay. Nice jacket. You might have an acceptance...thank you...or you might have a downplay...oh, this old thing. So, there are pairs of interactions and furthermore, these overlap. So, the question may be followed by an answer and the answer is then followed by another question.

A bit like bigram models or Markov models, it can also have sub-dialogues. "OK. There's two stops. Actually, what day of the week is the 15th? It's a Friday. Hmm, I would consider staying an extra day till Sunday. OK. OK. On Sunday I have--" So, this is a correction sub-dialogue. The extra bit changes the day and therefore, you go back to where you were before only with the day parameter changed.

Another sort of sub-dialogue is a clarification. "What do you have going to blah, blah, blah on the 5th? Let's see, going where on the 5th? Going to Hong Kong. Oh, here are some flights". So, this clarification, the system didn't understand where they were going, therefore asked for clarification and then returned to what do you have going to Hong Kong in this case. You can also have some pre-sequences.

The user actually wants to reserve a seat on the train to New York, but first of all he asked, "can you make train reservations?" And the system said, "yes, I can". So, this is before you actually ask the question, you're asking, is this a reasonable question to ask? So, there's also some extra complications to human interactions. Conversations are not typically just one person in charge, and the recipient is answering the questions.

There are special cases like reporters asking a chef or reporters interviewing someone.

But typically, most human conversations have mixed initiative. I lead, then you lead, then I lead. So, natural language processing systems are quite difficult to do this because they have to allow for the human to take charge and then for the system to take charge and so on. So, often they simply have either the user initiative system where the assumption is the user's going to ask a question, the system responds.

So, for example, IT help desk. The assumption is the IT user is going to want some information, so all we have to do is answer the user's questions. Or the alternative, of course, is to have a system issue. The system is asking questions, for example, to fill out a form. So, Hubert chatbot is trying to get students to give some feedback on the course and ask specific questions like, how could the course be improved, what did you think of the lecturer, and so on.

The users can't really to take over and start asking questions back. Both of these are not really conversations, they're either just answering the user's questions or just asking the user questions. A further complication which we won't even go into here is the system has to try to make inferences, like the agent says, "and what day in May did you want to travel?

The client says, "OK, I need to be there for a meeting that's from the 12th to the 15th". And from that, you have to infer that he has to travel before the 12th. It's no good traveling on the 12th because he won't be there for the meeting on the 12th. So, that's an inference the system has to make. The client hasn't explicitly said I need to travel before the 12th. And some of the challenges of human

conversations, in reality, human chatbots...human interactions are more complicated than chatbots. Chatbots can't really do this.

OK, so now let's have a look at some examples of real chatbots. So, ELIZA. You've probably heard of very famous first chatbot Weizenbaum back in 1960s developed this. It's an example of a transcript. Weizenbaum was a psychologist. He developed a chatbot which behaves in the style of a psychotherapist, and it's a particular sort of psychologist, a Rogerian psychologist. And the Rogerian approach is to engage the client in conversation and aim to get the client to understand their own problems by talking about their problems themselves.

The idea is to draw the patient out by reflecting the patient statements back at them, and this is good. This is good as a chatbot technique because you don't actually have to know anything about the client. You assume the pose of knowing that almost nothing about the real world. Whenever the client says anything, you just invite them to say more about it and then we don't have to know about what they're talking about. For example, the patient says, "I went for a long boat ride”.

The psychologist wants to get the patient to talk more about it, so you say, "tell me more about boats”. You don't assume what the boat...don't know what a boat is, you assume that the psychologist has some sort of goal, which is to get the patient to talk about things. So, there is a contest every year or so organised by Hugh Loebner called the Loebner prize contest, which is sort of a version of the Turing Test where several chatbots can enter and they're evaluated by humans to see if they think that they're human or not.

And the ones that win or the ones that try to win usually choose some sort of domain like this where they pretend they don't understand much about the world and they get the user to talk about it. So, how does this work? So, ELIZA is based on a large set of rules that Weizenbaum devised, and the rules are essentially some sort of pattern. And then the user input is matched against these patterns until it finds one pattern that it decides to use. And then if the pattern is matched, then the transformation rule which generates an output.

So, for example, zero you, zero me. Zero means any text. So, anything followed by you followed by anything followed by me is a pattern, and then the output or transform me is what makes you think I something you. The three in this case is the third word or the third constituent. So, do you like me. Like is the third one. What makes you think I like you is the response. "You hate me. What makes you think I hate you".

So, rules are organised by these keywords. Each keyword has a pattern and a list of possible transforms. So, the key word might be "you" and the pattern involving you is something, you, something, me and the possible responses, not just one but several of them, and then the ELIZA platform can choose one of those. And there's a sort of general structure for each keyword.

There may be a pattern followed by several transforms and there may be several patterns, each of which is followed by several transforms. OK, so that's a general set of lists of patterns. Then these keywords are then in order or ranked from a very specific one to a very general one. So, "I" is very general, so that would be down near the end. And the reason being that if the input doesn't match a very specific keyword, then it should match at least some of these things.

So, "I star" means any sentence starting with "I", you can come back with you say you something or other. "I know everybody laughed at me. You say you know everybody laughed at you". "Everybody" is more specific so you can have a more specific...if everybody followed by anything at all, then you can ask, who in particular are you thinking of? So, key words are stored with their rank. "Everybody" has a rank of five and "I" has a rank of zero.

And there's also a default set of responses. "Please go on", "that's very interesting", "I see", and so on. There's also this idea, as well as the patterns, of a memory. Whenever "my" is the highest keyword, then you would select a transform on the memory list and apply it to the sentence.

So, let's discuss further why your three and three is something on the memory list. So, it's possible...it's not just simply turn taking turn...we looked before as if every ELIZA reply was based on

what the user had just input, but you can also store away what the user had input earlier on in the memory so then you could go back to it.

This is another human-like thing that rather than just simply having a Markov or bigram model, you can, at various interesting points, maybe even at random, go back to an earlier thing. Earlier you said that, or does that have anything to do with what you said earlier on? So, if no key word matches a sentence...if basically the user inputs something where you can't find a rule that does...you can either try one of these very general things like tell me more, or you can go back to the memory phrase.

Then we see very simple rule-based architecture, and you can build your own chat box like this. I'll show you in a minute a platform you can try this out on. However, this a nice example where it's possible to implement an AI system and ELIZA it just went ahead and implemented it and tried it out without realizing some of the ethical implications. People became...at least some of them started to become emotionally involved with the program.

One of Weizenbaum's staff was using it and asked him to leave the room because she didn't want him listening. She thought this is personal and private. He thought he was just going to store all the ELIZA conversations for analysis, being a psychologist, and it would be interesting to explore.

But then people immediately pointed out, this is not ethical. There are privacy implications. You can't store people's conversations without their permissions. And this is a problem because you have to get their permissions in advance. You can't wait till afterwards because they thought they were having private conversations, even though they knew it was just a program. I suppose most ordinary people don't really have a sense of how the AI is not really intelligent, it's just a set of rules. That's ELIZA.

If you want to find out more, look in the textbook or even...yeah, just Google ELIZA. So, one more example of an extension to this. So, we've seen a very simple model patterns, for each pattern some responses, and there's some random element in choosing the responses. And if there's no pattern that matches, you have a memory so you can introduce things that we've talked about earlier on.

So, that brings in the idea of having, as well as the patterns, some further overall architecture, and that's the idea behind PARRY. This is also a clinical psychology focus, but this is used to study schizophrenia. This is a model of schizophrenia because in addition to the ELIZA sort of structure, it also has a mental state model. It has these global variables for anger, fear, and mistrust. These are internally just numbers, and they all start with low variable numbers.

And then after each turn, the user statement is analysed to see if it induces fear or anger. For example, if a user says an insult or if PARRY decides what the user said was an insult, that increases the anger. Or if a user says something nice, that decreases the anger and so on. But if the PARRY analysis says this is just an ordinary sentence, then the anger, fear, and mistrust can go down a bit.

So, PARRY's responses depend on the mental state, or rather the value of the variable, fear or anger. So, if fear was high, then you go for one sort of responses which involve running away. While if anger is high, then the responses are more hostile and so on. So, you can see you can have essentially variables in the chatbot which is separate from the patterns and actions and these variables have additional features.

And PARRY, in some sense, passed the Turing test in the 1970s. The Turing test is essentially a human has to judge or maybe some human judges if the output of the system is a chatbot or a real human. And in this case, psychiatrists were given transcripts of interviews with PARRY and interviews with ordinary people having paranoid schizophrenia and they weren't sure which was which. So, that did show that the PARRY transcripts looked reasonable.

OK. So, if you want to have a go at building a system like ELIZA, there is a...Pandora Bots is one example website where you can build your own or try some that are there. This is based on AIML, the artificial intelligence markup language. So, AIML is a bit like XML or HTML, that's the markup language for web pages, but it allows you to write the sort of patterns that we saw in ELIZA.

Richard Wallace, an AI researcher, developed a version of ELIZA he called ALICE and put it online so that anybody around the world could interact with it. And in addition, what he did...ALICE learned, but not in machine learning, but in rather Wallace learning. In other words, he monitored the interactions and if ALICE gave a poor or implausible response, Wallace added some extra rules to fix this so that next time around if a user asked the same thing it would have a more plausible response. So, he essentially built up a huge bank of patterns and responses over time. ALICE built a very large pattern list of plausible replies.

Now I mentioned the Loebner prize competition held every year where any number of different chatbots are entered and a number of humans, usually AI professors and other academics who know a bit about chatbots, they sit and interact with various systems and some of the systems are linked to chatbots. And some of the systems are linked to a person in another room and they're trying to judge which ones are human and which ones are not. And ALICE won the prize not just once, but three times for being the most plausible, largely because it was more plausible because it had more human interaction built into it.

OK, so that's ALICE. However, you don't necessarily have to have a huge set of rules if you're for a very limited domain. So, for some practical applications...and there are commercial chatbots for limited domains because that's all you need to talk about. For example, in the book by...the paper by Abu Shawar I recommend you read on "Chatbots, Are They Really Useful?" she gives several examples, one of which is FAQchat.

That is a frequently asked questions, and FAQchat can be trained with any set of frequently asked questions. There's a frequently asked questions for Python website and you can feed this into the FAQchat chatbot and then if you ask any questions, it will try to find the question in the FAQ which is most similar and give you the answer from the FAQ, so that's fairly straightforward.

And if you ask, who is the president of the United States, then it will find the Python question most close to that and give you the answer. It will be the wrong answer, but then you shouldn't have asked

a silly question. OK? So, within the limited application of asking questions about Python, that is a quite sensible approach.

OK, another example is when a chatbot is asking questions in a very specific type of domain. Remember I said that in real conversations, both partners take turns in leading the conversation. For chatbots, the simplifying assumption is either the user is asking questions and chatbot has to answer them, as in FAQchat, or the system is asking questions and the user simply has to answer them.

And this is, for example, in student feedback interviews, Hubert asks, did you like course, what could you improve in the course, and so on. Or in job applicant interviews, the interviewer is in charge and the applicant simply has to answer the questions. That's what Hubert does. And there's a recommended paper on University student surveys using chatbots about our project here at Leeds University on using Hubert.

OK, so there's some examples of chatbot systems that are rule based. Now going to see a couple of corpus-based chatbots and corpus-based chatbots are particularly good for general conversations. So, if you have a very specific domain, then you can build a set of rules from an FAQ website or from some other standard source.

You go on a very general system, then you may have a big corpus and whatever you do as a question or input is, it has to match. It has to search the corpus for something similar and that generates the response. You can do this by using information retrieval to grab a response from the corpus or possibly a bit more sophisticated is use the corpus to train a language model, maybe some sort of encoder-decoder type neural network so that when the user typed something in, the trained language model, which is essentially trained on the corpus, will generate a response.

And so, the first one is memory based learning or instant space learning. So, the retrieval says whatever the user input is, you search through the corpus to find the nearest match. The alternative is you use the corpus in advance to train a very large neural network, so the neural network essentially

encodes all of the corpus so that when the user types an input, the generated output is appropriate for the corpus.

So, where did you get these corpuses from? Well, only computational linguists, also corpus linguists such as myself. We had collected corpora of language, which include not just written text but also spoken. So, the British National Corpus, 100 million words of British English, includes 10 million words of spoken texts transcribed and 19 million words of written text in computer readable form.

And also, the international Corpus of English or ICE, that has for British, American, Australian, New Zealand, Canadian, and various other variants of English, a million words including a significant portion of spoken dialect transcribed, and it's not just English.

So, for example, the paper on "Chatbots, Are They Really Useful?" mentions the Afrikaans chatbot, which was trained on the Korpus Gesproke Afrikaans, a corpus collection collected at Potchefstroom University in South Africa of Afrikaners speaking to each other. And it then, having trained on that, it was able to converse in the style of the Afrikaners.

OK. Another way of doing this, quite popular in America, has been to pay people to talk to each other on the phone or even to give them free phone calls. So, transcripts of telephone conversations between volunteers, so the switchboard or call home corpora. So, call home means we will give you a free one-hour call to your parents or family back in wherever home is from America.

So, you get lots of transcripts of, for example, Arabic, where Arabic speakers in America call home to their family in Saudi Arabia or Tunisia or wherever it is. And apart from that, there's lots of questions and answers data sets like the frequently asked questions or at Leeds University, the IT help desk has a very large data set of typical IT user questions and their answers, and this can be used to train a chatbot.

OK, there are other sources. So, there's a huge corpus of movie dialogues because there's lots of corpora of movie subtitles. Most movies nowadays, apart from the actual speech, they have subtitles added and you can extract the subtitles and that gives you the transcript of what was said.

You can, if you want to, just simply hire people to have conversations. That's a bit like the call home thing, but it's usually for if you want to have a corpus on a very specific topic. So, if you want to have conversations about, let's say, how to get into university, then one way is to hire some people to talk about how to get into university.

There are also pseudo conversations. If you go onto social media, you can look at interaction on Twitter where people respond to other ones or Reddit or Chinese Weibo and so on. These are noisy in the sense they're not really conversations, so they can be used as a sort of pretraining to get some sort of language model for the neural network type approach.

One ethical issue for all of this is you must remove personally identifiable information. People will just not...obviously people's own names may not be mentioned, but they may mention other people's names or even their telephone numbers or their credit card numbers or all sorts of things that anchor telephone conversations, so you do have to be careful about that.

OK, so how does this work? If you've got a large corpus, if a user says a Q, question, and it has got training corpus, then you find in the corpus the turn that is most similar to the question using something like tf-isf cosine measure and then you respond with R.

Or perhaps a bit more complicatedly, having found something similar to the user question you say the response to R, assuming that the corpus is essentially a series of questions and answers. So, the user asks a question, you find the question in the corpus or something similar to it, the most similar to it, and then whatever the answer was, you give that. And that's a very simple model and it seems to work.

The neural method isn't that much more complicated. You give a user a question...given a user question, you find the turn in the corpus which is most similar. The difference is here you use BERT rather than using cosine similarity. BURT, we'll see later on what that is. That's basically a huge neural network for noticing if two sentences or similar or not. And then you just say response, or you say the response to the question.

One final way you could use a neural network for is if you have a whole of a neural network encoding the corpus. In other words, you train the corpus for every sentence, every question and response. You put this into the neural network, so the neural network essentially builds a model of, for any input, what is the appropriate output. Then when the user typed in a question, a query, the response is generated by conditioning the encoding of the query and the response so far.

So, you basically generate a response using the neural network. It's, as far as I can see, just a fancy way of doing what we've already done but using a neural network instead. And just as an aside that there is a slight problem with generating something which replies to whatever was said before because you can get into loops, if you like.

If anybody here who...if you've heard of PG Wodehouse wrote a series of books about Jeeves very early on, about a hundred years ago. Here's a quote. "'What ho!' I said". 'What ho!' said Monty. 'What ho! What ho!' 'What ho! What ho! What ho!' After that it seemed rather difficult to go on with the conversation". So, what ho is a standard response to anything but if you then reply with "what ho", then you're stuck.

And here's some real examples. "Where are you going?" "I'm going to the restroom". "See you later". "See you later". Don't say see you later in response to see you later because if you do, you're stuck. OK, so that's a slight problem. One extra thing apart from having a corpus is that to generate responses, you can also use a corpus to find informative text rather than simply a response to a dialogue.

So, Xiaoice, if you ask, tell me something about Beijing, it won't simply look in its corpus to find what word sounds like tell me something about Beijing and then respond with this next question...the next sentence. It actually looks into news articles or public lectures or things like Wikipedia and finds some information. So, that's another use of a corpus.

OK. So, finally, chatbots are fun and they're good for narrow scriptable applications, things like answering questions about the IT help desk. On the other hand, you have to remember that they're not really people. They don't really understand, and the problem is they do give the appearance of understanding, so that may fool some people that may get involved with them. And they break down.

The rule-based chatbots in particular can take a lot of learning lots of rules and they still, unexpectedly, don't give really silly answers. The information retrieval-based chatbots are as good as their training data. So, if there's something bad in the training data, they can give bad responses. So, maybe you want to integrate this chatbot into some sort of knowledge-based frame-based agents, and that's what the idea of a dialogue state or belief state architecture is.

So, a more sophisticated has to have dialogue acts, and most modern systems don't just have a corpus trained chatbot or a set of frames, but some sort of combination. And this is working into these industrial systems. Things like Amazon Alexa probably have this sort of slot understanding as well as the chatbot understanding. So, here we have an architecture diagram.

A typical user says something, there's automatic speech recognition, and then spoken language understanding which feeds whatever is said into a dialogue state tracker. And the dialogue state has a corpus of what sorts of things are likely to be responses, but also a state model of frames, like you want to go from downtown to airport and so on. It may have some sort of dialogue policy model, which says what's an appropriate thing to say back, and then it generates some natural language response as a text and then the text replies by text to speech.

That's a more sophisticated model overall. It's natural language understanding which extracts it. It doesn't just do transcription of speech to text, but also extracts the slot fillers from the speaker. And

there's a dialogue state tracker which maintains the current state of the dialogue and some sort of dialogue policy where the system decides what to do next rather than simply replying with a best match. And the natural language generation module produces more natural utterances rather than simply whatever is in the corpus. So, each of these parts is a bit more sophisticated.

So, here's an example. User says, "I'm looking for a cheaper restaurant”. And the system works out the important thing is cheap. The price has to be cheap. So, the system says, "sure, what kind and where?" And the user replies, "Thai food, somewhere downtown”. That means the system now knows is looking for cheap, he's looking for Thai food, and he's looking for downtown or centre of the city. So, the system finds something which matches and gives that as an answer. "The House serves cheap Thai food”.

The user comes back with, "where is it?" That means the system now needs to know that he, as well as wanting cheap Thai...price cheap, food Thai, and area centre, he is also asking for an address, therefore comes back with an answer which matches all of that. Not just any old address, but an address which matches the previous answer, The House. OK. So, we see this more sophisticated dialogue tracking model.

OK. Finally, I want to look at how to evaluate dialogue systems and some ethical issues in dialogue systems. So, evaluating is quite hard. You used to, for classifiers, coming up with an accuracy score or maybe a precision and a recall score. For dialogue systems you can't really say, is the response correct or not correct 100% in a conversation. If I ask you something, there isn't only one correct answer and everything else is just wrong. There are usually several plausible answers, several variations of answers. So, if it's a task-based dialog like trying to book a ticket, then you can measure the task performance. At the end, have they managed to book a ticket?

If it's a conversational chatbot, am I having a reasonable chat with you? And this is down to human evaluation. You have to say, overall, did I like the conversation? Did I feel I was having a reasonable chat? And that's much more subjective. But in the whole, chatbots are evaluated by humans. The

humans talk to a chatbot and either decide, in the Turing test, was it human or not, or maybe give them a score on a rank.

You also have observer evaluation, a third party who watches the conversation or maybe reads a transcript of a conversation and assigns score. As in the example of PARRY, PARRY had a conversation with users and then the psychologists were asked to rate whether or not that PARRY was having a conversation or a real psychopath or psychotic person...sorry...was having a conversation.

So, in both cases is it it's by human evaluation. In the first case, it's the human user is getting to feel how good their talk was and the other one is to have a bit more objective, a third party. Maybe a conversation expert and AI professor could be doing this. And you can do this in even more sophisticatedly. Rather than just saying was it good or not, you can have a sort of very specified experiment.

For example, human will chat with models for six turns and then rate on eight dimensions of quality. Did they avoid repetition? Was it interesting? Did it make sense? Was it fluent? Was the system listening to me? Was it inquisitive? Was it really human-like? Was it engaging? All these sorts of issues. And each of these, in the evaluation, particularly for having an objective third person doing the evaluation, they may have quite detailed definitions.

So, rather than just avoiding were they repetitive, they have criteria like repeating themselves over and over or sometimes said the same thing twice or always says something new. So, that's three different levels of repetitiveness. So, for each of these different dimensions or features, you rate on a scale of one to three or one to five. So, you see that this can get quite sophisticated, and you end up with a score...not just is it good or not, but a more nuanced score for each chatbot.

So, typically and the most objective way of doing it, the most scientific way of doing it, is to have people have conversations with chatbots and then have annotators...for example, AI professors or students...analyse the responses, look at the conversations, and typically to make it easier, you have

two conversations and decide which one is better. So, you might have a conversation with a real human and a conversation with a chatbot and decide which one is which. That's one way of doing it.

And according to all of these criteria, engagingness, which would you prefer to talk to in a long conversation? It's a bit subjective, but you can see these are the sorts of things which you decide...it's not just is this a real person, but this is a natural person I'd like to be able to talk to? Automatic evaluation is still difficult. We saw from machine translation...machine translation is still pretty much judged by translators deciding this good or not.

There are these blurred scores for machine translation, but there's no real equivalent for chatbot evaluation. There is a research direction called adversarial evaluation if you train a Turing-like classifier to distinguish between human responses and machine responses. So, you can actually build a machine learning classifier trained on examples of chatbots being positive...sorry…examples of chat box as against examples of human conversations and then given a new unknown conversation, it tries to decide whether or not it is chatbot or human. And the more successful a dialogue system is at fooling this classifier, the better of a system. It's actually quite hard because we haven't got lots and lots of training data and we're not quite sure it works that way, but you can imagine that's how it might work.

The other thing is, as in any IT system, if you want to evaluate it you can have a user satisfaction survey. So, you can set something like Hubert to ask the user after a chatbot had...after they've had a conversation you can ask, for example, was the system easy to understand? Did the system understand what you said? So, you're actually asking in the same way as you might ask, did you like the graphical user interface? You can ask the user, did you like the chatbot?

They know that the system was a chatbot. We're not trying to say, did it fool you into thinking it was a human, but simply, how good was it to use? How easy was it to use? Do you think you'd use this system in future? And this is much more common for practical chatbots for specific purposes, like the IT help desk chatbot or some of the example chatbots you'll see on the videos.

Other things which aren't really measuring the humanness of a chatbot at all, but things like efficiency cost. How long did it take? Or quality cost, how many times the system had to say I don't understand you, number of times the user had to barge in. So, these are more or less systems engineering evaluation or IT systems evaluation metrics applied to chatbots. These aren't really chatbot specific evaluation measures.

So, that's how to evaluate. The answer is we're not very sure, but we can evaluate them in the same sorts of ways as you evaluate other IT systems and IT interfaces. Finally, I want to look at some ethical issues if you're using an IT system, then you have to...as an interface, then a general way of building interfaces is to study the users and the task and build some sort of simulations.

A Wizard of Oz study is based on a famous story of Oz where the wizard of Oz appeared to be an all-powerful wizard who knew many things, but actually turned out to be just an American salesman. So, the salesman behind it was just an ordinary person. So, in the same way, a Wizard of Oz study is having a real person pretending to be the chatbot and entering into conversations.

And then you have conversation transcripts between human and pseudo chatbot to give you an idea of what the chatbot should be talking like, so you set iterative tests on users. And one issue is, particularly if you're having a Wizard of Oz study, then the users get used to the idea that it is actually very human-like because there really is a human there. So, distinguishing between humans and chatbots becomes important. In artificial agents you have to make clear that you're not simply building it to see if it's possible to build.

An old example before there were computers, Mary Shelley wrote this book called Frankenstein where she created a human out of bits of other humans, sewing them together without consideration of any ethical issues. So, just the idea was the doctor wanted to find out if it was possible to take bits of both human bodies and sew them together to make a human again. And of course, when they did it, when they succeeded, the human was very upset because they didn't feel normal, didn't feel right, and lots of things happened in the story which weren't very good.

So, there is a problem with a chatbot. There are various ethical issues. There are certainly safety issues, for example. If a system starts distracting the driver while they're driving or if it's a medical chatbot and it starts giving bad medical advice or if a system is being used by psychology as psychotherapy for mental health patients and the system starts being nasty or abusing the users, then they could be put in danger.

There's also if a system starts being biased or nasty, not just about the user, but about particular social groups, then that's representational harm. And of course, there's also this problem of information leakage, that the chatbot, if it's trained on a corpus, it knows what's in the corpus. It may know things about the people, what they said in the corpus, which may leak out. If somebody in the training corpus read out their Social Security number, then the system may say their Social Security number back to the user.

OK. So, as I said before, it's very important if you're using a chatbot for mental health but you don't say anything, it's going to be bad for the mental health, or for in-vehicle conversational agents. If you come to Leeds University, we actually have a driving simulator, which is a real car where they've taken out the actual mechanics and put in computer graphics and computer-generated movement, so the car moves around and stuff...so it feels like it's moving around anyway.

And this is used for things like testing out interfaces to different computer control systems. And when I've tried it, I have been aware that it is a virtual system and I've also been very aware that this chat thing talking to me is distracting me from driving so that I crash into things. Luckily, they're only virtual lorries that I crash into. So, you have to be aware of the environment and the driver’s level of attention in developing the chatbot.

Here's a real example. Microsoft developed a Twitter chatbot, and they gave it the personality of a young 18 to 24-year-old American woman. They trained it on lots and lots of Twitter data from such people and it was allowed to tell jokes and asked people to send selfies.

It used informal language and slang and emojis from the training corpus. They found a training corpus. And it was designed to learn from the users so people who used it, their interactions were added to the corpus. However very soon, Tay became offensive and abusive, and it started giving out conspiracy theories and Nazi propaganda and harassing women and reflecting racism and misogyny, and that's because it was reflecting what was already in Twitter.

The problem is that there's no filter on Twitter. Anybody can post anything on Twitter and people were posting misogynistic, racist, and inflammatory tweets, and therefore it was learning them and spouting them back. So, Microsoft had to take Tay down after only 16 hours. So, you have to think about...in the design phase you can't simply assume that whatever the users say should be included in your system. You have to have more filters about this.

And this pointed out the issue that actually all trained data sets, if you take real data, is likely to have biases in them. So, they try and take applying hate speech and bias detectors on training data sets for dialogue systems. Twitter, Reddit, and other dialogue data sets have already said how useful for training data, but they found lots of bias and hate speech in the training data. And therefore, because it's in the training data, the dialogue models trained on this also include hate speech and bias.

A quite standard example, the British National corpus, 100 million words, was developed by a number of researchers. One of the researchers in this made quite a name for himself by writing a series of research papers pointing out or built on the essential thing that, OK, the most frequent words in the British National corpus are function words like the, often, and.

The most frequent content word which isn't a function word is actually a F-U-C-K or fuck used as a noun or verb or an adjective. And that's because in real British English that happens to be a very frequently used word. However, if you have a conversational agent or any other chatbot, if it keeps on using that word that may be unacceptable to certain applications.

OK. Another issue is privacy in the corpus. There may be things like computer, turn on the lights. And then if you answer the phone, you may say, hello, yes, my password is...or yeah, the user may be out buying something and therefore gives out his credit card numbers. That happens a lot in real corpus.

So, I've listened to transcripts of telephone conversations which include these sorts of things. Oh, sorry. I've not listened to; I've read transcripts of conversations that actually include these sorts of things. There may also be intentional information leakage. Dialogue systems designed to send user data or dialogue systems are designed to ask the user to give them their credit card number. If you want to buy something you have to give your credit card number.

So, it's important that if you're going to use this training data then you have to preserve the privacy in the training by removing it or blanking it out in some way. OK, so that's some of the ethical issues in chatbots. So, to summarise then, we've had an introduction chatbots and dialogue systems. We've seen that chatbots are...you might think of them as just basically the user says something, a chatbot replies, or a user ask question, the chatbot gives the answer.

Human conversations are much more complicated. There are turn taking like question and answer, but there are other sorts of turns. There's dialogue acts. There's also things like interruptions. You have to decide when to stop, when to give way and let the other person have a turn. There are essentially two general sorts of architectures, rule-based systems or machine learning systems.

The rule-based systems humans have devised long sets of rules, and we've seen some examples of this. Or the corpus-based chatbot, you get a large dialogue corpus and you learn from that in one way, either by instance based learning...basically having all the dialogue act as being examples...or by a neural network, building a model off the corpus and then generating the response from that model.

Nowadays current commercial systems like Alexa have some sort of combination of corpus-based modelling and also state based modelling, so you have a dialogue state architecture. For chatbots, there's a real issue in evaluating them. Its usually humans are involved in some way. And as with

other text analytics systems, we always have to be aware of ethical issues. Chatbots seem to be human-like and interacting with chatbots we take it for granted that they're going to be human-like.

But really, we also have to be careful that they don't start spouting Nazi propaganda or giving out personal information that they shouldn't do. OK. For more information, please read the textbook chapters on question answering systems and on chatbot and dialogue systems in the Jurafsky and Martin textbook. Also read some of the papers that are recommended because there is some useful background on ALICE and Hubert and AIML and so on. OK. Thank you very much for listening and I hope you enjoyed the lecture. Bye for now.