
Chapter 12: Machines Who Talk

Introduction

In 1979 Pamela McCorduck published 'Machines Who Think', a survey of the then-nascent field of Artificial Intelligence (it has recently been re-issued with an afterword [1]). Apart from the shock of the sentient pronoun, McCorduck's book also helped raise the hype-level of AI in the 1980s. The field has tended to disappoint since - technologies such as expert systems promised much but seemed to vanish into smaller-than-expected niches. Natural Language Understanding systems were part of that early wave, with expected applications in automated assistants, innovative computer interfaces, machine translation and police and military security. The dream has never died, but the applications subsequently seemed somehow subscale against the promise.

Why is this of interest for next-generation networks? There is an astonishing disparity between the two types of traffic which NGNs carry: application data traffic vs. audio/video calls. When you point your browser at an application transactional website, you have an HTML-based 'conversation' with the application where the syntax, semantics and pragmatics is completely specified. This allows complete automation of each step of the transaction and arbitrary amounts of machine intelligence and formatting can be applied. However, when you use the NGN to talk to another person, to download music or publish or observe video content, the network can see the media stream (as a byte stream) but the intelligence it can bring to bear is normally only at the signal layer itself. Typically this is limited to compression for transport efficiency. What a fantastic opportunity for new services and revenues if the next-generation network could actually understand and produce conversation and video!

As with most advanced technologies, we are not there yet, but neither are we out of the game completely. I will start by looking at automated conversational systems currently in service. I will then look at the most straightforward approach to designing such systems, the so called 'chatbots', and analyse both how they work and why they are dead-ends in their current form. Next I will look at what has to be done to put together an effective conversational system and some of the reasons why this is hard. Finally I will outline some of the prospects and their likely impact on the NGN upper layers.

The state of the art

I recently had lunch with Andy MacLeod, former CEO of what is now Verizon Europe. I asked him what he thought the hottest issues would be in telecoms over the next few years.

“Going back to my materials science roots, I have to say new kinds of batteries based, for example, on direct methanol fuel cells, which could last maybe two weeks. You go into a shop and buy a top-up, just like with cigarette lighters - that could transform portable devices.

“My next call would be WiFi. It may be cheap and cheerful, but, like Ethernet, it just keeps on improving. With the new mesh architectures allowing scalable municipal networks, both the technology and the economics is looking better and better for new Internet-based access networks. I think WiMAX is really going to have a hard time given WiFi’s market momentum.

“For my final candidate, recall the Star Trek movie where they go back in time to present-day earth. The Engineering Officer needs to use a current-day computer to design something. He’s in the offices of the engineering company and he reaches over to a laptop, picks up the mouse and says ‘hello computer?’ Pretty amusing. I have some experience in the area, back when I used to run one of Nortel’s speech processing businesses. It would be fantastic to have a system which did what the Enterprise’s engineering guy wanted. I don’t know if we’re near a breakthrough which would get us there, though.”

There is a name for the kind of system Andy wants to see, and it’s called a Spoken Dialogue System.

Traci

This morning, I picked up the phone to call a hotel reservation service. Normally, you are placed in a queue, waiting for a human agent to take the details of your request, but Premier Hotels has an automated agent called *Traci* (T).

[Ringing tone]

T: “Hello, I am Traci, Premier Hotel’s automated reservations assistant. Do you have a member’s profile?”

Me: “No.”

T: “That’s OK, I can help you make a reservation or check availability. You can interrupt me at any time, and if you need help, say “help” and if you need to speak to someone, say “agent”. So, which premier hotel are you enquiring about?”

Me: *“The Andover hotel.”*

T: *“I think you want ‘Dover Central’. Is that correct?”*

Me: *“No.”*

[Note that this is the hardest part of the dialogue for the automated speech recognition system to handle. The chain has hundreds of hotels, many of which sound similar. ‘Andover’ must have been recognised as ‘An’ (ignored) + ‘Dover’ and ‘Hotel’ must have been mapped to ‘Central’. What the system hasn’t communicated at this point is that it is expecting just the name of the hotel or town with no other extraneous words. It doesn’t have a problem with ‘No’ though. So now I get coached.]

T: *“Please say only the name of the hotel you are enquiring about or say the name of the town or city where you want to stay. You can also say ‘help’ ...”*

Me: (interrupting). *“Andover.”*

T: *“I think you want Newbury Central, is that correct?”*

[This mistake is slightly mysterious as the sounds are very different. However, Andover is a very new hotel, and Newbury is close to the town of Andover, so it may have been a database look-up error.]

Me: *“No.”*

T: *“Please say only the name of the hotel you are interested in staying in, ...”*

Me: (interrupting) *“Andover.”*

T: *“What’s the arrival date?”*

[Success!]

Me: *“Tomorrow.”*

[The system has no problem with indexicals - relative references, to 'date' here.]

T: "*For how many nights?*"

Me: "*Two.*"

T: "*How many adults, aged 16 or over, will be staying?*"

Me: "*One.*"

T: "*And how many children, 15 or under?*"

Me: "*None.*"

[Notice that this is entirely a *Traci*-centric conversation. I only have to utter single word replies from a highly-restricted vocabulary. As a consequence, my answers are correctly recognised and the 'conversation' runs smoothly.]

T: "*Thank you, so you'd like ... [summarises the order] ... Is that correct?*"

Me: "*Yes.*"

T: "*OK, please wait while I check availability ... [Confirms available room and price] ... Which would you like, make the booking, check the availability of a hotel nearby, or change this enquiry?*"

Me: "*Change this enquiry.*"

T: "*So which premier hotel are you enquiring about, or you can just say the name of the town or city?*"

Me: "*Do you take pets?*"

T: "*I'm sorry, I didn't understand what you said. Could you please say only the name of the hotel you require ...*"

[Notice how brittle the conversation is. Any deviation from the built-in script causes the transaction to crash and burn. At this point, the only recourse is to talk to a human being. However, it is unlikely that the ordinary caller at this point will remember that to do so, they have to say ‘agent’ as they were instructed in the first conversation turn. Most people are likely to experience a rapid increase in frustration at this point and hang-up].

[Hangs up].

Traci works just fine for routine and predictable transactions, and otherwise fails ungracefully. Once callers are themselves trained on *Traci*'s limitations, then I am sure they prefer to deal with ‘her’ than to wait for five or ten minutes in a queue. It’s really the audio equivalent of filling in a form on a web page. And the economics of software instantiations replacing human agents explains why *Traci* was waiting for my call in the first place.

Traci can talk about very little other than making hotel reservations for the Premier company. At the other end of the spectrum there are systems which purport to be able to talk about anything at all.

The Loebner prize

Dr. Hugh Loebner set up this competition in 1991 as an application of the Turing test for artificial intelligence. Each year a number of programs compete to appear most human in their responses before a panel of judges (the judges are typically psychology professors). The interaction is like instant messaging: each judge types a message to the ‘chatbot’ in a pane on the screen, and back comes a reply.

How convincing is the state-of-the art in chatbots? Well, the 2005 winner was a program called Jabberwacky. This program is on the Internet (at <http://www.jabberwacky.com/>) and anyone can ‘converse’ with it. Here is my attempt (I am ‘user’).

Jabberwacky: Do you live in Chicago?

User: I live in England.

Jabberwacky: Is england full of programmers?

User: It's normally full of rain.

Jabberwacky: Well what are you up to?

User: I'm writing a book.

Jabberwacky: About e-mail addresses?

User: **You seem obsessed with computer topics!**

Jabberwacky: Bots and maze theory in data structures, are our next upgrade in tech.

User: **How did you win the Loebner prize?**

Jabberwacky: No. But then you didn't either!

It looks like Loebner will be waiting a while to award his \$25,000 prize to a system which judges are incapable of distinguishing from a human being!

Why are these systems so poor? They work by superficial manipulation and rearrangement of the text the user types in, plus a few stock, pre-loaded responses. The idea is that with enough templates and canned phrases, something approaching 'normal' human responses can be obtained. Typical rules might be:

Rule n:

```
(user): Do you <verb> <object>?
=> (answer): Yes, I <verb> <object> most days.
```

So the user types "Do you like ponies" and the chatbot disturbingly replies "Yes, I like ponies most days". It would have worked better if the activity had been 'eat ice-cream'! Presumptively-trivial activities like sorting out verb endings are handled by pre- and post-processing stages. The chatbot mechanism also manages user-specific state information. So it can ask for the user's name, store it, and regurgitate it later as in this rule.

Rule m:

```
(answer): What is your name?
=> (user): <text> -> $name
=> (answer): Hi, $name, what do you do?
```

The chatbot sends "What is your name?". You type in "Peter" (or "My name is Peter" and the first three words are stripped off) and the word 'Peter' gets bound to the variable '\$name'. The chatbot then answers "Hi, Peter, what do you do?" Convincing to some, perhaps?

The first program like this was Eliza, written by Joseph Weizenbaum in 1966 in the style of a non-directional psychotherapist. It succeeded in fooling many users with its sympathetic responses, prompting a certain degree of horror and disillusion on Weizenbaum's part about the human condition [2]. However, chatbots are of little use in external-goal-directed activity because they contain little knowledge about anything in the world, and have less ability to do anything with that knowledge. At their best, they hold a mirror to the user. For more information, check the *Personality Forge*, a site dedicated to helping people design and run their own chatbots at www.personalityforge.com. It has contributed a number of prize winning systems.

Putting chatbots to one side, it is time to return to systems currently in service. How do they work, what do they do, and how can they be made to perform better?

Spoken Dialogue Systems - the state of the art

The army is in a far-away land, fighting a vicious but diffuse insurgency. The prisons and requisitioned barracks are full to brimming with locals, the results of innumerable sweeps through slums and shanty towns. Most of the captives are probably innocent, but how do you tell? Hardly anyone in the army speaks the local language.

In a high-tech army, most problems are believed susceptible to technology. The army uses an interrogation program called GSTAPO1 (General Speech Translation And Production Operation system mark 1). It has been refined in the field, and here is how it works in practice.

The suspect is brought into the interrogation room and strapped to a steel chair. He or she faces a table, bolted to which is a heavy duty microphone. Speakers and video cameras are visible high on the walls, while the floor is solid concrete striped with cracked and stained guttering. After a scene-setting announcement from the speakers, designed to encourage the suspect to cooperate, a pleasantly insistent synthesized voice poses a number of benign, unthreatening questions.

Q1. "What day is it today?"

The suspect often would not know, because they had been incarcerated for days, but the system corrects them and asks them again.

Q1a. "It's Tuesday. Now, what day is it today?"

Q2. “What year is it?”

Q3. “What is the name of your capital city?”

Q4. “What is your favourite sport?”

Q5. ...

The intent is to calm the prisoner and get them into a routine of cooperative responses. Some prisoners simply refuse to cooperate, or hurl abuse at the microphone. Fortunately, the chair comes with some persuasive technology which could be used to provide encouragement to respond in such cases. Since there is no time limit to sessions, there are significant opportunities for Pavlovian conditioning.

Once the suspect is cooperative, they are taken through a protocol scientifically designed to complete an optimized interrogation profile: name, address, occupation, family details, religious and political affiliation, personal history, recent activities and so on. This is cross-referenced in the database with collateral information, and certain responses are ‘trigger’ items which automatically mark the prisoner’s status as more or less interesting. Note that GSTAPO1 is conversing with captives in their own language, but completing the interrogation profile in a language that military intelligence people can understand, thus addressing the critical linguistic barrier.

GSTAPO1 is an effective first-stage filter, and allows the great majority of non-insurgents to be released straight back into the community. However, it is inadequate in two regards. Firstly, there is the problem of type 2 errors, ‘false negatives’. These are the bad people who learn how to ‘game’ the system and pretend innocence: GSTAPO1 mistakenly flags them for release. However, setting the thresholds for a presumption of innocence to a very high level simply means that few can pass it. The result is a mass of ‘problematically-bad’ guys who cannot be released and who overwhelm the small number of skilled interrogators fluent in the local language. Secondly, the people marked as likely insurgents need a more sophisticated interrogation style than the ‘tick-in-the-box’ approach of the current system. Again, if there were enough human interrogators, fine. But there are not. So the military put in a request for GSTAPO2.

To understand why GSTAPO1 was of limited utility, and how its shortcomings might be fixed, we have to dive inside its internals a little. GSTAPO1 (figure 1) is a traditional, state-of-the-art spoken dialogue system.

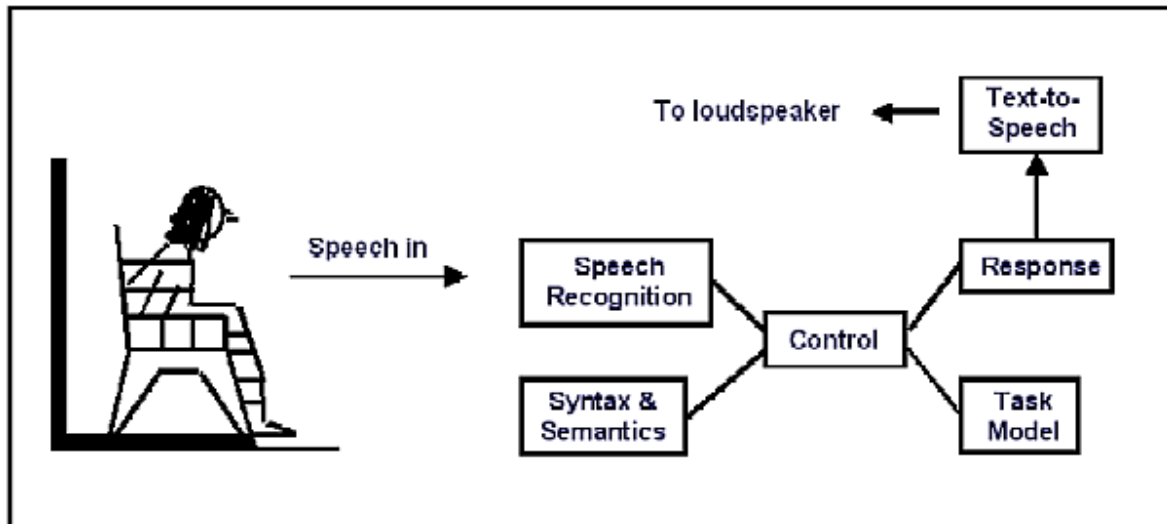


Figure 1. The architecture of GSTAPO1

The speech recognition module is a standard commercial chipset and statistical package which picks up the sequence of phonemes directed at the microphone by the subject, and matches them against phoneme sequences which correspond to words. Words are not always pronounced in a standard way, unfortunately. Reasons for variation in what is heard at the microphone include:

- background noise, coughing
- regional accents
- personal speech idiosyncrasies
- stress and fatigue
- variability in the time taken to speak the word
- mis-starts and hesitations
- variant pronunciation of the same word in different contexts
- age, gender of speaker.

For these reasons, a straight look-up of ‘what was heard’ in the phoneme-to-word database throws up many possible matches. The next stage of processing, syntax and semantics, narrows these down. The first technique is statistical. By analyzing large numbers of interrogations, it is evident that certain pairs of words have a significant probability of adjacency, whilst other combinations are seldom heard. For example

Likely	Unlikely
threw grenade	few grade
ceiling collapsed	scene claps
no water	nor adder

Table 1. Likely and unlikely word combinations

In practice, we don't try to model the whole of language, just the collection of words which are likely to be relevant to the task in hand, namely primary interrogation. GSTAPO1 uses this kind of statistical model and also something called a contextual grammar. This is best illustrated by an example. Suppose the prisoner is asked "where were you born?" There are a number of ways he might reply (supposing he was born in a place called Barin and was inclined to be truthful):

Q. "Where were you born?"

- A1. "Barin"
- A2. "I was born in Barin"
- A3. "<expletive> Barin <expletive>"
- A4. "Barin, Barin"
- A5. "Uh .. er ... it was Barin in 1983"

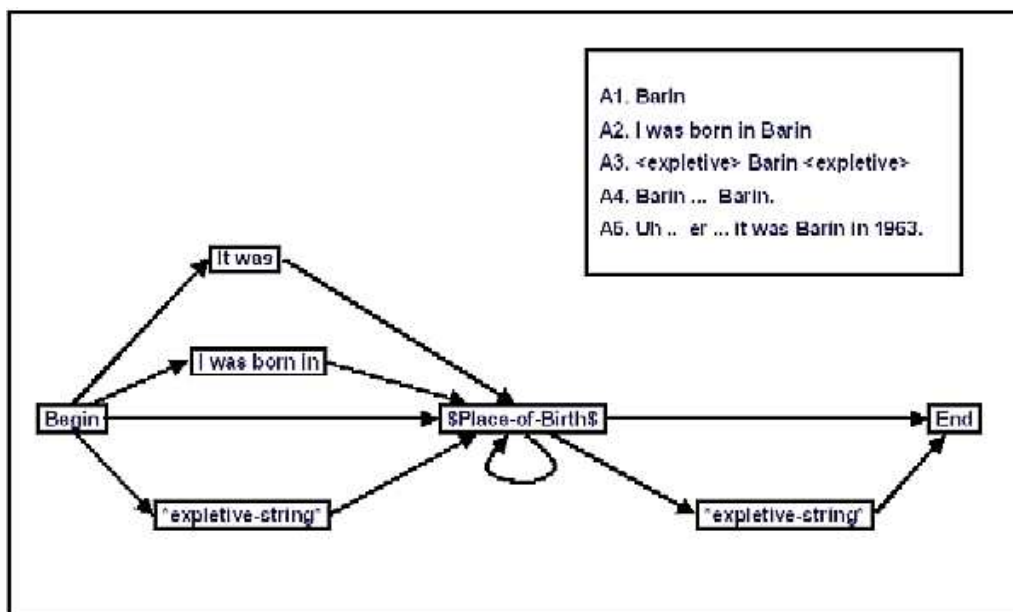


Figure 2. Grammar for 'place of birth'

Assume, based on many pre-recorded interrogations, that this is a good-coverage set of responses. How do we pick out the information we need? The answer is via a grammar network as shown in figure 2 (notice that the grammar will also pick-up related responses not on the above list, such as any answers which end with an expletive).

As the reply from the prisoner is acoustically processed, the speech recogniser is looking, bottom-up, for words which match the sound signal, and which statistically are likely to go together. At the same time, the syntax and semantics component is using the grammar of figure 2 to identify which path through the network the prisoner is taking. If there is a way to get the task model variable \$Place-of-Birth\$ bound to a recognised word (and the system's dictionary will identify words which are place names), then the system will reply:

Q. Confirm you were born in \$Place-of-Birth\$ - answer 'yes' or 'no'.

If the answer is 'no', or the attempt to understand the previous question was inconclusive, the system will re-ask the original question:

Q. "Where were you born?"

The final module of GSTAPO1 is 'control'. This structures the overall dialogue. More sophisticated systems keep a track of what they have learned and what they still need to find out, and ask the next question based on some prioritization of what is, as yet, unknown. However, the present system operates a pedestrian pre-planned dialogue model - another network, part of which is shown in figure 3.

The dialogue control module starts once the suspect has been made ready. It is a canned speech which simply instructs him or her what is to come, and how they are to behave. There then follows the *calming dialogue*, a sequence of benign questions of no intelligence value, but which allow a certain amount of speaker training in the recognition software, and speaker training in the sense of getting the prisoner to a state where he/she is prepared to answer questions and the system can understand the responses. Then the system gets down to business.

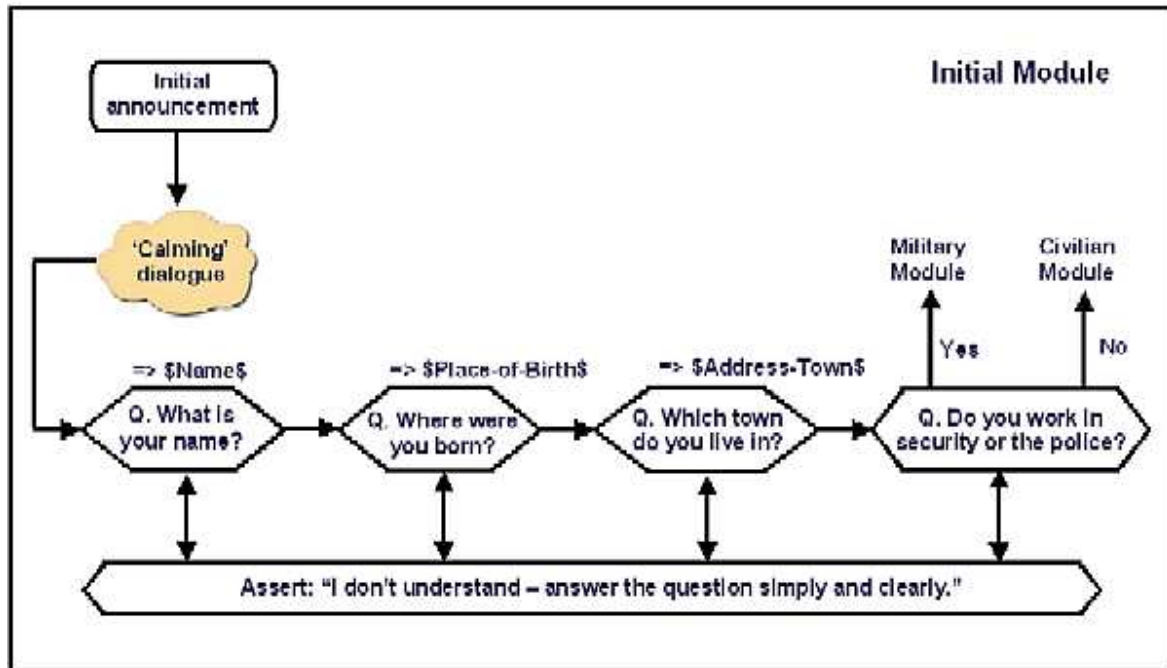


Figure 3. Part of the overall dialogue control module

First there are the standard questions: name, birth-place, address. Each of these questions has its own Q. and A., a contextual grammar as previously shown in figure 2. Successfully-recognized responses allow the task model to be updated - the interrogation profile for the prisoner - and lead to further progress within figure 3. If the system cannot understand a response, control passes to the lower box (“I don’t understand – answer the question simply and clearly.”) and the question is re-asked.

From the prisoner’s point of view, the system is unbearably pedantic, ignoring volunteered information, checking everything with a follow-up ‘yes/no’ question and taking forever to do the debrief. In intelligence terms, this is a plus, as boredom and repetition helps the debrief process. More commercial systems attempt to accept additional information if it is offered, combine confirmation with further questions and handle a wider range of conversational gambits as in this example.

Q. “Do you want a large or extra-large burger with fries?”

A. “Extra large and can I have it with extra sauce too?”

Notice that ‘extra large’ refers to the burger, and that the ‘it’ is ambiguous without knowing more about burger-bar practice and the menu.

Spoken Dialogue Systems - raising the game

The military would like to automate dialogues like the following

Q1. "Where were you on Thursday evening?"

A1. "I was at home."

Q2. "You were not. You were seen in the old town working on a truck. What were you doing?"

A2. "Did you say Thursday?"

Q3. "We know you were there. <X> has told us everything."

....

Why can't GSTAPO1 handle this kind of dialogue? Because this conversation isn't pre-planned, it's more like a chess game between two players (but with an open-ended set of pieces!). To play you have to know a lot about how things work in the world (places, times, travel, trucks, bombs, ...) and a lot about motivations, why people do things. You also need to make an accurate assessment of what point the other person is trying to make at each stage of the conversation (their move, if you like) so you can find the right conversational countermove with a view to getting an admission. What, for example, is the interrogating party meant to make of this response:

A2. "Did you say Thursday?"

And even the standard problems still exist, waiting to trip the system over. It's easy to say that we won't worry too much about sophisticated syntactic processing of utterances, because people don't speak grammatically anyway. True, but then someone answers like this.

Q. "Who entered the base?"

A. "The men from the militia with the bombs."

Noting as an aside that we got back a noun phrase, not a sentence (ellipsis), what does it mean? Figure 4 shows two parses of the phrase, one with the men belonging to a militia which had bombs, the other having the men from the militia themselves having the bombs. We can't tell which is meant, but unless we know there is ambiguity, we can't understand what was just said, and we can't ask the right follow-up question. So it seems we do have to build-in quite a bit of grammar knowledge.

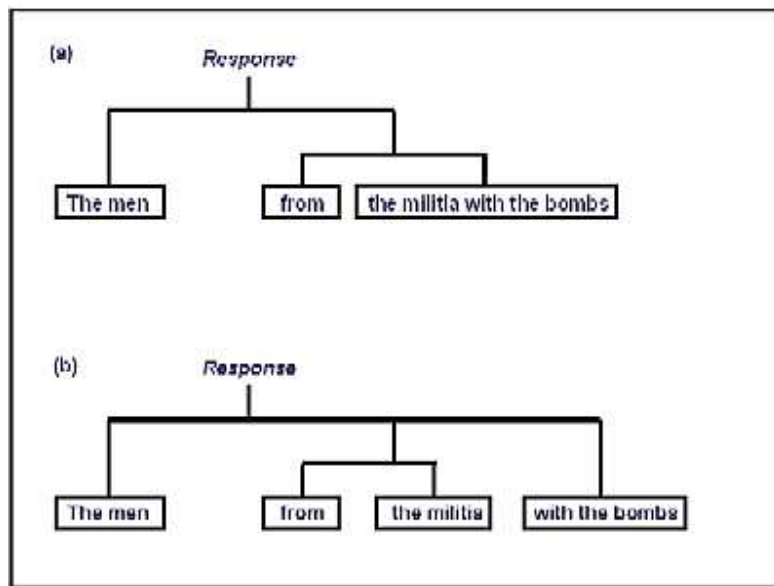


Figure 4. Ambiguous syntax

When the military asked for the development of GSTAPO2 to undertake this kind of sophisticated mixed-initiative dialogue, military R&D told them that what they were requesting was impossible. It is beyond the state of the art to:

- deploy such a wide competency in language, grammar and meaning
- build in the encyclopedic level of real-world and situation-specific knowledge required
- model and infer the suspect's beliefs, desires and intentions accurately
- understand the nature and structure of dialogue itself, sufficient both to understand the real import of what is said and to construct effective replies
- exhibit expert-level interrogation skills.

The R&D officer wearily informed his audience that the system they were asking for would not only be able to pass the Turing test with ease, but would also be an effective military interrogator! This combination of tasks would defeat most *people*, let alone present-day Spoken Dialogue Systems.

GSTAPO1 is fictitious (as far as I know), but is based on the architecture for spoken dialogue systems described in [3] chapter 4. The issues with GSTAPO2 are discussed in more detail in [4], especially chapters 4, 35 and 37 (p. 672). Professor Pulman, mentioned next, drew my attention to this site: <http://language.cnri.reston.va.us/TeamTIDES.html> where the military are working on technologies which would make systems like this possible.

The view from Oxford University

For a clearer view of where we are with language understanding, I discussed the current state of the art in conversational systems with Professor Stephen Pulman, Professor of General Linguistics at Somerville College, Oxford University. Stephen was enthusiastic about recent developments in computational linguistics and this was part of our discussion.

...

“We have an enormous amount of textual information available on the web, and very powerful syntax resources and processing engines. We can look at how certain nouns or verbs are used in context in websites across the Internet, and begin to classify relationships such as ‘is-a’ and ‘part-of’ in a meaning hierarchy using statistical clustering. From a few starter examples, such as ‘Canada is a country’, ‘Spain is a country’, there are systems which can automatically fill-out the relationship, generalizing to other countries.

“This may not sound too exciting, but, for example, I expect soon to be able to type into a search engine something like ‘find me a good price on an Epiphone’ ... “

“What’s an Epiphone?”

“... they make guitars - and the system understands the concept of ‘guitar-making company’ and can perhaps also suggest something from Fender or Gibson. I expect question-answering to be another area where some new competencies will be on display quite soon.”

“Like ‘Ask Jeeves’? Those kind of systems, as I recall, were pretty hit and miss - really just keyword search.”

“And they still are, but once a search engine has built a complex hierarchy of linked concepts, then new kinds of more intelligent search suddenly become possible. You could type ‘When was President Nixon elected?’, assuming you wanted to know, and actually get the answer, rather than a list of websites which happened to include those keywords, which is what happens today.”

“What about other kinds of applications, for example, the automation of call centre agents?”

“Well, I don’t know the quality of your conversations with service staff over the phone, but it seems to me to be anything but straightforward usually. The conversations always seem plagued by mis-statements, misunderstandings and perhaps some emotion too.”

“If we can’t get it right with people, I suppose there is no hope for automated systems.”

“Well, the limited telephone bandwidth doesn’t help. You’d be surprised the difference it might make to add a video connection. Once we can make a video call to the call-centre, then the far end can see our face and our lip movements - even our gestures. Research has shown that this extra information can substantially improve accuracy.”

“Are there any other applications you can see coming along?”

“You know, I used to be surprised by how few applications there were for genuinely interactive natural language systems. I used to think this was due to the tedium of communication through a keyboard ...”

“I guess instant messaging and texting might be counter-examples to that?”

“... perhaps, but I now think the reasons might be rather deeper. Somehow there needs to be a sense of talking with a real person, a presence. Perhaps we need to link the spoken dialogue systems with household robots to create an ‘embodied agency’, make it real.”

“You mean an artificial person, or a child?”

“It’s not totally far-fetched, there’s an EU research project looking at precisely that. It’s called CoSy - Cognitive Systems for Cognitive Assistants, and it’s looking at integrating many sub-disciplines within AI to put together a robot capable of acting and communicating with understanding, perhaps indeed at the level of a small child [5].”

“Sounds expensive and difficult.”

“Maybe so, but I suspect there is a real market for AI-based assistants and conversational partners, particularly with an aging population. It’s an interesting question how much better we have to get than the current generation of chatbots so that an embodied conversation agent would be a genuine boon to people

needing support. After all, most dialogue is about maintaining social relations rather than answering questions or solving problems. This is an area where we really need to resolve some tough issues to figure out how to do it well enough to be effective in practice!”

Conclusions

Putting the scientific questions of theoretical linguistics to one side, the practical engineering of conversational systems has had some successes.

- Call-centre agent-replacement systems are in service today, although restricted to fixed-dialogue standardized functions such as booking flights and hotel rooms.
- Chat-bots with a wide but superficial language skill have achieved some level of dialogue competence, and there are some business models struggling to get launched, for example, in language-learning practice.
- Dictation systems have found a market, and after user-training have achieved astonishing accuracy levels.

Current research is leveraging extensive banks of lexical, syntactic and concept-organization material available over the Internet to induce large-scale concept hierarchies. These will find their use in making search engines more powerful and supporting new kinds of queries based on knowledge and inference.

Some of the bottlenecks to achieving full ‘Turing test competence’ include the lack of progress in understanding how to capture the meaning of conversation, and the difficulties of understanding exactly what is involved in participating in human dialogues. Is progress here dependent on having embodied systems available which humans can interact with as part of an extended social grouping?

Looking ahead, it seems likely that progress will be more focused on niches where conditions are susceptible to rapid progress, rather than some across-the-board advance to a new paradigm of human-system interaction. But there again, research has a habit of being unpredictable.

References

- [1]. McCorduck, P., *Machines Who Think*, AK Peters, Ltd.; 2nd edition, 2004.

- [2]. Weizenbaum, J., *Computer power and human reason: From judgment to calculation*, Penguin, 1984.

- [3]. McTear, M. F., *Spoken Dialogue Technology*, Springer-Verlag, 2004.

- [4]. Mitkov, R., (Ed.), *The Oxford Handbook of Computational Linguistics*, Oxford University Press, 2004.

- [5]. CoSy project website. <http://www.cognitivesystems.org/>.

Recommended Reading

Many aspects of speech understanding and dialogue systems are currently making rapid progress, driven by the existence of more powerful computers, on-line Internet tools such as corpora, grammars and knowledge bases, and demand pull from human-computer interface, search engine and general Internet applications communities.

Mitkov's 786 page compendium [4] contains 38 relatively compact chapters aimed at newcomers to the field (although assuming a mathematical/computer science background). It offers a remarkable coverage, with many pointers to further reading and research.

McTear's book on Spoken Dialogue Systems [3] is much more focused on systems of this type. He is able to get into practical details of how such systems are structured and implemented, providing a strong reality check for those who might believe that science-fiction systems such as 2001's HAL are just around the corner. Why is it hard? McTear will tell you.