Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone

Tuesday, May 8, 2018

Posted by Yaniv Leviathan, Principal Engineer and Yossi Matias, Vice President, Engineering, Google

A long-standing goal of human-computer interaction has been to enable people to have a natural conversation with computers, as they would with each other. In recent years, we have witnessed a revolution in the ability of computers to understand and to generate natural speech, especially with the application of deep neural networks (e.g., Google voice search, WaveNet). Still, even with today's state of the art systems, it is often frustrating having to talk to stilted computerized voices that don't understand natural language. In particular, automated phone systems are still struggling to recognize simple words and commands. They don't engage in a conversation flow and force the caller to adjust to the system instead of the system adjusting to the caller.

Today we announce Google Duplex, a new technology for conducting natural conversations to carry out "real world" tasks over the phone. The technology is directed towards completing specific tasks, such as scheduling certain types of appointments. For such tasks, the system makes the conversational experience as natural as possible, allowing people to speak normally, like they would to another person, without having to adapt to a machine.

One of the key research insights was to constrain Duplex to closed domains, which are narrow enough to explore extensively. Duplex can only carry out natural conversations after being deeply trained in such domains. It cannot carry out general conversations.

Here are examples of Duplex making phone calls (using different voices):

Duplex scheduling a hair salon appointment: -00:00

Duplex calling a restaurant: -00:00

While sounding natural, these and other examples are conversations between a fully automatic computer system and real businesses.

The Google Duplex technology is built to sound natural, to make the conversation experience comfortable. It's important to us that users and businesses have a good experience with this service, and transparency is a key part of that. We want to be clear about the intent of the call so businesses understand the context. We'll be experimenting with the right approach over the coming months.

Conducting Natural Conversations

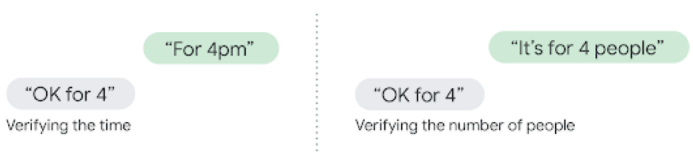
There are several challenges in conducting natural conversations: natural language is hard to understand, natural behavior is tricky to model, latency expectations require fast processing, and generating natural sounding speech, with the appropriate intonations, is difficult.

When people talk to each other, they use more complex sentences than when talking to computers. They often correct themselves mid-sentence, are more verbose than necessary, or omit words and rely on context instead; they also express a wide range of intents, sometimes in the same sentence, e.g., *"So umm Tuesday through Thursday we are open 11 to 2, and then reopen 4 to 9, and then Friday, Saturday, Sunday we... or Friday, Saturday we're open 11 to 9 and then Sunday we're open 1 to 9."*

Example of complex statement: -00:00

In natural spontaneous speech people talk faster and less clearly than they do when they speak to a machine, so speech recognition is harder and we see higher word error rates. The problem is aggravated during phone calls, which often have loud background noises and sound quality issues.

In longer conversations, the same sentence can have very different meanings depending on context. For example, when booking reservations "Ok for 4" can mean the time of the reservation or the number of people. Often the relevant context might be several sentences back, a problem that gets compounded by the increased word error rate in phone calls.

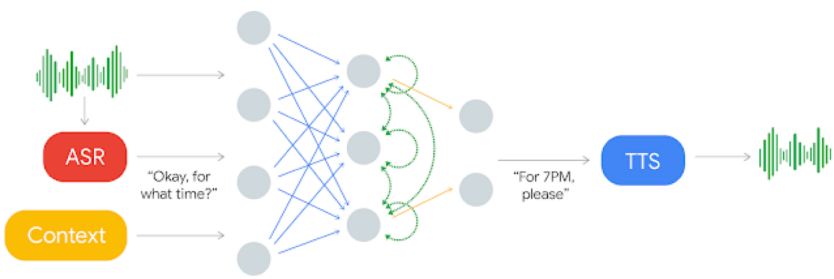


Deciding what to say is a function of both the task and the state of the conversation. In addition, there are some common practices in natural conversations — implicit protocols that include *elaborations* ("for next Friday" "for when?" "for Friday next week, the 18th."), *syncs* ("can you hear me?"), *interruptions* ("the number is 212-" "sorry can you start over?"), and *pauses* ("can you hold? [pause] thank you!" different meaning for a pause of 1 second vs 2 minutes).

Enter Duplex

Google Duplex's conversations sound natural thanks to advances in *understanding*, *interacting*, *timing*, and *speaking*.

At the core of Duplex is a recurrent neural network (RNN) designed to cope with these challenges, built using TensorFlow Extended (TFX). To obtain its high precision, we trained Duplex's RNN on a corpus of anonymized phone conversation data. The network uses the output of Google's automatic speech recognition (ASR) technology, as well as features from the audio, the history of the conversation, the parameters of the conversation (e.g. the desired service for an appointment, or the current time of day) and more. We trained our understanding model separately for each task, but leveraged the shared corpus across tasks. Finally, we used hyperparameter optimization from TFX to further improve the model.



Incoming sound is processed through an ASR system. This produces text that is analyzed with context data and other inputs to produce a response text that is read aloud through the TTS system.

Duplex handling interruptions: -00:00

Duplex elaborating: -00:00

Duplex responding to a sync: -00:00

Sounding Natural

We use a combination of a concatenative text to speech (TTS) engine and a synthesis TTS engine (using Tacotron and WaveNet) to control intonation depending on the circumstance.

The system also sounds more natural thanks to the incorporation of speech disfluencies (e.g. "hmm"s and "uh"s). These are added when combining widely differing sound units in the concatenative TTS or adding synthetic waits, which allows the system to signal in a natural way that it is still processing. (This is what people often do when they are gathering their thoughts.) In user studies, we found that conversations using these disfluencies sound more familiar and natural.

Also, it's important for *latency* to match people's expectations. For example, after people say something simple, e.g., "hello?", they expect an instant response, and are more sensitive to latency. When we detect that low latency is required, we use faster, low-confidence models (e.g. speech recognition or endpointing). In extreme cases, we don't even wait for our RNN, and instead use faster approximations (usually coupled with more hesitant responses, as a person would do if they didn't fully understand their counterpart). This allows us to have less than 100ms of response latency in these situations. Interestingly, in some situations, we found it was actually helpful to introduce *more* latency to make the conversation feel more natural — for example, when replying to a really complex sentence.

System Operation

The Google Duplex system is capable of carrying out sophisticated conversations and it completes the majority of its tasks *fully autonomously*, without human involvement. The system has a self-monitoring capability, which allows it to recognize the tasks it cannot complete autonomously (e.g., scheduling an unusually complex appointment). In these cases, it signals to a human operator, who can complete the task.

To train the system in a new domain, we use *real-time supervised training*. This is comparable to the training practices of many disciplines, where an instructor supervises a student as they are doing their job, providing guidance as needed, and making sure that the task is performed at the instructor's level of quality. In the Duplex system, experienced operators act as the instructors. By monitoring the system as it makes phone calls in a new domain, they can affect the behavior of the system in real time as needed. This continues until the system performs at the desired quality level, at which point the supervision stops and the system can make calls autonomously.

Benefits for Businesses and Users

Businesses that rely on appointment bookings supported by Duplex, and are not yet powered by online systems, can benefit from Duplex by allowing customers to book through the Google Assistant without having to change any day-to-day practices or train employees. Using Duplex could also reduce no-shows to appointments by reminding customers about their upcoming appointments in a way that allows easy cancellation or rescheduling.

Duplex calling a restaurant: -00:00

In another example, customers often call businesses to inquire about information that is not available online such as hours of operation during a holiday. Duplex can call the business to inquire about open hours and make the information available online with Google, reducing the number of such calls businesses receive, while at the same time, making the information more accessible to everyone. Businesses can operate as they always have, there's no learning curve or changes to make to benefit from this technology.

Duplex asking for holiday hours: -00:00

For users, Google Duplex is making supported tasks easier. Instead of making a phone call, the user simply interacts with the Google Assistant, and the call happens completely in the background without any user involvement.

A user asks the Google Assistant for an appointment, which the Assistant then schedules by having Duplex call the business.

Another benefit for users is that Duplex enables delegated communication with service providers in an asynchronous way, e.g., requesting reservations during off-hours, or with limited connectivity. It can also help address accessibility and language barriers, e.g., allowing hearing-impaired users, or users who don't speak the local language, to carry out tasks over the phone.

This summer, we'll start testing the Duplex technology within the Google Assistant, to help users make restaurant reservations, schedule hair salon appointments, and get holiday hours over the phone.

Yaniv Leviathan, Google Duplex lead, and Matan Kalman, engineering manager on the project, enjoying a meal booked through a call from Duplex.

Duplex calling to book the above meal: -00:00

Allowing people to interact with technology as naturally as they interact with each other has been a long standing promise. Google Duplex takes a step in this direction, making interaction with technology via natural conversation a reality in specific scenarios. We hope that these technology advances will ultimately contribute to a meaningful improvement in people's experience in day-to-day interactions with computers.