Trevor Donnell
December 7, 2001
6.191 Preliminary Thesis Proposal

Hands-Free Internet using Speech Recognition

Introduction

The hands-free Internet will be a system whereby a user has the ability to access websites using his or her voice rather than having to use a keyboard to enter the name of the site. This system will have the potential for retrieval of headlines, stock quotes, traffic reports, sports scores, weather information, email, etc. The system will function in a similar way to a hands-free cell phone in that it will allow users to say the name of the website they wish to visit. The system will be capable of reading back the aforementioned features of the Internet, such as headlines, etc, via a speech synthesis system capable of putting together sentences if given the words that should be in the sentence. The system should work in a way that makes it sound conversational, as if the user was speaking to an actual human rather than a computer.

Technology

The technology for this system was developed by the Spoken Language Systems (SLS) Group in the Laboratory for Computer Science at MIT. The SLS Group has created various core technologies to perform speech recognition (SUMMIT), natural language understanding (TINA), discourse and dialogue modeling, language generation (GENESIS), and speech synthesis (ENVOICE). The following sections will describe each of these systems in further detail, with examples of how each of them works in the MERCURY air travel service.

The example sentence for the examples that follow is “Is there a flight from Boston to San Francisco Friday?” (taken from the SLS website).

Speech Recognition System (SUMMIT)

SUMMIT is the speech recognition system developed by the SLS Group. Spoken language creates certain acoustic signals and patterns that SUMMIT can convert into a sequence of distinct words. The basic system used in SUMMIT matches the acoustic signals to signals stored in a database, so that the system can deduce first the phonemes and then the words that the user is trying to convey. SUMMIT uses a probabilistic language model to make a decision about exactly which sentence it believes the user has just said. For example, using the sentence above, the following list is generated:

1. Is there a flight from Boston to San Francisco Friday?
2. Is there a flight from Austin to San Francisco Friday?
3. Is there flight from Boston to San Francisco Friday?
4. ...

Natural Language Understanding System (TINA)

TINA is the natural language understanding system created by the SLS Group. It takes a sentence from SUMMIT as input and parses the sentence in order to figure out its meaning. TINA is embedded with many semantic models of common grammatical structures, so it has a basis for splitting a sentence into its basic components, such as subject, verb, object, predicate, etc. These models are used in order to add semantic tags to the grammar so that TINA can extract a semantic representation of a sentence (called a semantic frame, shown below). In this way, TINA can logically determine what the user wants based on what he says by finding the deeper semantic meaning of the sentence.

```
Clause: EXIST
        Topic: FLIGHT
        Quantifier: INDEF
                Predicate: SOURCE
                Topic: CITY
                        Name: Boston
                Predicate: DESTINATION
                Topic: CITY
                        Name: San Francisco
                Predicate: TIME
                Topic: DATE
                        Day: Friday
```

Dialogue Modeling

In order to be able to perform a user's request, the dialogue manager must determine the relevance and completeness of the request, gather the information requested from a database and create a semantic frame for the response. The system has a built-in set of rules with which the user must comply, and the system can request additional information from the user if necessary. When the user has provided the system with enough relevant information pertaining to the request, the dialogue manager retrieves the desired response from the database and creates a semantic frame similar to the following (again based on the Boston to San Francisco sentence):

```
Clause: AVAILABILITY
        Flights found: 3
                List:
                Topic: FLIGHT
                        Date: October 19
                        Airline: United
                        Flight number: 163
                        Departure Airport: BOS
                        Departure Time: 7:00 AM
                        Arrival Airport: SFO
                        Arrival Time: 10:23 AM
                        Stops: 0
                Topic: FLIGHT
                        Date: October 19
                        Airline: United
                        Flight number: 161
                        Departure Airport: BOS
                        Departure Time: 9:00 AM
                        Arrival Airport: SFO
                        Arrival Time: 12:22 PM
                        Stops: 0
                Topic: FLIGHT
                        Date: October 19
                        Airline: American
                        Flight number: 195
                        Departure Airport: BOS
                        Departure Time: 9:00 AM
                        Arrival Airport: SFO
                        Arrival Time: 12:37 PM
                        Stops: 0
```

Language Generation (GENESIS)

GENESIS is the system which creates a natural language representation of a semantic frame, in essence reversing the effects of TINA. The dialogue modeling system inputs a semantic frame to GENESIS, and GENESIS converts the frame into a standard response in one of several languages such as English, Spanish, Mandarin, Japanese, or even HTML or SQL. Using the above example, the following sentence is created:

> I have 3 nonstop flights:
> A United flight arriving at 10:23 AM,
> a United flight arriving at 12:22 PM,
> and an American flight arriving at 12:37 PM.
> Please select one of these flights or change
> any constraint you have already specified.

Speech Synthesis (ENVOICE)

ENVOICE is a concatenative speech synthesis system developed by SLS. In other words, it creates speech by piecing together segments of speech which are prerecorded in a database. This concatenation can take place at the phrase, word, or syllable levels depending on the needs of the sentence and the existence of various sounds in the database. The system makes the speech sound like human speech by adjusting the pitch or overall flow of each syllable so that the breaks between various concatenated sounds is as smooth and unnoticeable as possible. In this way, the speech created by ENVOICE sounds much like the speech created by a real human, thus adding to the conversational aspect of the overall system.

Hands-Free Internet

The hands-free Internet system will provide users with the option of accessing websites without having to use a keyboard or mouse for input. The user will speak into a microphone the address of the website he or she wishes to retrieve, and the system will retrieve that website. The user may then request headlines if it is a news site such as CNN.com, weather reports if it is a weather site such as weather.com, sports scores at sites such as ESPN.com, stock quotes from money.msn.com, and more. The hands-free

Internet system will use each of the technologies described above in certain ways which will each be discussed shortly.

First of all, the system needs to figure out which website the user would like based on his input. An example would be "I would like to visit ESPN.com." SUMMIT will take this sentence and build up word recognition based on the different sounds the user makes as the sentence is said. The basic SUMMIT system will need to be extended with the capability to recognize the names of websites and distinguish them from words with the same sounds, and one way this could be done is to enhance SUMMIT with the ability to recognize the endings in the website names, such as .com, .net, .org, etc. Then the system would be able to deduce that the input directly before the ending was, in fact, a website.

In the second step, SUMMIT passes the now fully recognized sentence to TINA, so that the natural language processing can occur. In the case of a website, TINA will create a semantic frame that resembles the following:

```
CLAUSE: Retrieval
        TOPIC: Visit
                PREDICATE: Website
                URL: ESPN.com
```

TINA will need to be extended so that it may recognize when a word passed to it by SUMMIT is a website rather than being a proper name or city name. Again, the endings on the names of websites tend to point out to users that it is a website, so the system can create the semantic frame requesting the website.

The third step will require the dialogue modeler to figure out that the user requested a website by investigating the semantic frame created by TINA. Again, this part should not be difficult except that the dialogue modeler needs to be extended to recognize the tag URL and then the name of the website. It should then go to the Internet, retrieve the website, and pass the HTML content on to GENESIS. The dialogue modeler should know from the name of the website what type of content appears on the site. For example, ESPN.com should alert the system to a sports site.

GENESIS should then create a sentence that says to the user something similar to the following: “Website is ready, what scores would you like to know about?”

ENVOICE will then create the sounds necessary to produce the sentence in synthesized speech, so that the user knows without looking that the system has retrieved ESPN.com and is ready for the next input, which may be something such as: “Did the Bruins win last night?” The process then starts over and determines that the user wants the score of the Bruins game that took place last night and whether the Bruins’ goal total was higher than that of the opponent.

Problems

Unfortunately, the system as described in this paper will never actually see the light of day. This is due to the fact that the SLS group simply does not have the resources to tackle a problem as large as this. Even just putting one of the functions into the system, such as retrieving headlines, is a massive project in itself beyond the scope of an M.Eng thesis. This all came to light in a meeting with T.J. Hazen, who described the goals of the SLS group. They, in fact, want to contain within their system a model of every possible input a user could ever give the system, so that the system can retrieve the correct action from the database and perform said action.

One major problem lies in the fact that the Internet does not have a set semantic scheme for the websites contained on it. Therefore, a single website is free to change its format as often as it likes without having to tell anyone what is going on. Unfortunately for this project, the World Wide Web Consortium is only now beginning to work on such a semantic standard for websites, so this project is totally infeasible at least for several more years. Only after a standard for website content comes about can this project theoretically be done.

A second problem T.J. Hazen mentioned is the fact that all websites contain at least some graphics and more often these graphics contain important text, such as the most important headline of the day. Without having a program capable of retrieving text from JPEG or GIF images, some of the desired content is lost from the website.

It is unfortunate that the project will not happen in the near future with the SLS group, but at least this comes at a time when it is still possible to find a project which can actually be done. Another unfortunate circumstance was the fact that this project only was rejected this week, because Dr. Hazen had been away from MIT for a couple weeks before the meeting. That is why this paper covers the project, simply because there was no time to find another one this week.