The citation information:

The Making of Task-Specific Chatbots for English Conversation Practice: A Preliminary Proposal

Shuen-shing Lee
National Chung Hsing University

Abstract:

The paper proposes tentative guidelines for the making of speech-to-speech chatbots for ESL conversation practice. The paper first presents a brief assessment of iOS Siri’s strengths and its possible application in the making of an ESL chatbot. In Section 2, the paper suggests that, to heighten the chatbot’s voice recognition rate, its soundbase should be built to match user speech input based on non-native-speaker-specific, area-specific, age-group-specific, and gender-specific parameters. Also noted is the working relation between an ESL chatbot and its clouding systems used in gathering area-specific accents. In Section 3, the paper suggests that a chatbot’s built-in replies be level-specific, topic-specific, and scenario-based to heighten its semantic recognition rate. In addition, the chatting avatar in a scenario should take the initiative in asking questions and use a particular soft-tone approach. The last sections of the paper deal with randomness in a chatbot’s responding behavior. As an alternative to randomness, the paper recommends that a Rogerian therapist mode of inquiry be adopted to augment the chance of the user’s continuing to engage in the dialogue.

Keywords: chatbot, Siri, speech recognition, semantic recognition, Rogerian psychotherapy

I. Introduction: Repurposing Siri

Aside from a human speaker with fluent command of English, a speech-to-speech (STS) chatbot is the most ideal partner for an ESL student in conversation practice. But why is a proficient ESL chatbot so hard to find in the marketplace? The answer lies in the fact that speech recognition technology has made very few breakthroughs since its inception in 1940s, compared to advances in other fields of digital technology. This explains why, at this writing, text-to-text (TTT)
and text-to-speech (TTS) chatbots run rampant on the internet while their STS counterparts are rare. Against this backdrop, iPhone 4S’s “Siri,” a “personal assistant” equipped with STS capacities, stood out upon its debut in October 2011. Its official promotional trailer (Thisismynextvideo, 2011) shows dialogues impeccably conducted between the user and Siri the chatbot. In this movie clip, daily routines such as emailing, time-setting, and scheduling, are all perfectly accomplished by voice exchange. In the early days after Siri’s release, user tests on the accuracy of its speech recognition have proven only slightly flawed. For example, in darkesseofmagyk’s movie clip, 11 questions, diverse in subject and mostly inquiries beyond the domain of daily routines, are presented to Siri and the chatbot scores 10 out of 11, based on the accuracy rate of its responses in terms of speech and semantic recognition. The one response of Siri’s which logically goes awry is that to a “weird” question: “I need to hide a body.” We will come back later to the question and the semantic cognition issues this question raises. Presently, what concerns us is speech recognition. Siri’s high competence in recognizing speech in general questions is unprecedented in the history of STS chatbots. This recognition accuracy has made the engineering of successful ESL chatbots much simpler in technical terms. With that in mind, we believe that an ESL chatbot empowered by Siri’s strength or its kind, like that of the Android “Assistant,” can handle ESL conversation practice satisfactorily, mostly because the ESL chatbot is task-specific, rather than for general purposes.

II. Suggestions for the Improvement of Speech Recognition Rate

In our discussion below, unless specified otherwise, we use as a point of reference an imaginary ESL chatbot meant for a Taiwanese 9th grader. The chatbot comes as an interactive companion to an ESL textbook used by the student. He/She is expected to have a vocabulary of the 1200 English words compiled by Taiwan’s Ministry of Education for the ninth year of English learning. In this view, the chatbot is content-specific. Note that learning materials in the textbook also include audio and video formats, which will help prepare the student to comprehend the chatbot’s speech in conversation practice.

In constructing the soundbase of the ESL chatbot, the following reminders can help heighten its speech recognition rate: the soundbase should be built to match user speech input based on non-native-speaker-specific, area-specific, age-group-specific, and gender-specific parameters:

1. A non-native speaker’s intonation usually differs greatly from a native speaker’s and, as such, sounds for each English word recorded and built into
the soundbase should include variations that may be encountered from non-native speakers. This will enable the chatbot to recognize and respond to utterances that may deviate from standard English.

2. Pronunciation by non-native English speakers from discrete areas, countries, or cultures differs from each other greatly. As such, it is advisable that a chatbot’s soundbase be area-specific. For example, a huge disparity exists between Taiwanese and Cantonese accents in spoken English.

3. Because the human voice changes with age, a chatbot maker should be sure that the sound spectrum considers the target user’s voice range.

4. Gender-specific differences in intonation should be taken into account since a boy’s voice differs from a girl’s.

Two more parameters will be added to the list, but first a bit of backstory on Siri is in order. Less than one year after its official launch, Siri has been criticized for getting poorer in both speech and semantic recognition, contravening Apple’s claim that Siri is improving with time its accuracy rate in both realms (Nivas, 2012). Steve Wozniak, Apple’s cofounder, remarked that he used to get the exact answer for the question, “What are the five biggest lakes in California?” but now in answer to the question Siri gave him information on realty listings (Samson, 2012). Another user, Test-Pilot, reports:

When it [Siri] first came out, I couldn't sing its [sic] virtues enough. It was great; I had very rarely experienced difficulty in getting Siri to do exactly as I asked. But lately, it’s getting whacky. Example: "Remind me to put the platter into the oven at 6:00 PM" gets transcribed as "Put Clara into the oven at 6:00 PM". Try again, it gives me: "Put Schlatter into the oven at 6:00 PM". Once more: "Put Hannah into the oven at 6:00 PM". There was one or two more before I gave up and typed it in manually. (Test Pilot, 2012; Italics mine)

In Wozniak’s case, Siri is “getting dumber” in semantic recognition. One of the technical reasons is that servers with search engines, which Siri is connected to, have been constantly gathering information from the internet and updating their storage to such a point that Siri has too much data to regurgitate. In the early days, Siri, using the same recognition algorithms, was able to identify a much more precise answer from a shorter list of options. Now the list has grown into a long queue beyond Siri’s semantic recognition ability. Kalinich (2012) fittingly points out: “The issue here is that Apple has shifted Siri to more of a universal search feature. This means that if
you ask a question it is going to hit Apple, then it might hit Google or another search engine other than Wolfram Alpha,” the particular server(s) Siri used when it was a stand-alone application. Kalinich continues, “You can still get the answer, but you have to pre-text the question with ‘Wolfram Alpha’.”

Siri is always learning, using clouding systems to assemble accents from its North American user group. The advantage is that it can quickly archive a huge array of accents but one consequence is that the archive will soon blow up, fraught with too much aural ambiguity. These ambiguous sounds have flooded the soundbases, breaching the maximum range of recognition tolerated by each individual standard sound. As such, Siri is no more able to persistently recognize regular sounds such as “platter,” as observed in Test Pilot’s report, and can answer the question only in a random way. That is why “Clara,” “Schlatter,” or a less relevant “Hanana,” popped up in the response.

By accepting into their servers all speech input without discrimination, Apple seems to intend to transform Siri from a task(routine)-specific chatbot into a general-purpose one. The failure, as illustrated by the two examples above, testifies to two things: 1. the making of a general-purpose chatbot currently remains an elusive goal; 2. general-purpose AI has a long way to go. Be that as it may, the original Siri can make a successful ESL chatbot, if its strength is appropriately directed. A little shrinking twist of its task in compliance with the above four suggestions would enable Siri to function satisfactorily for ESL learning.

Clouding systems of Siri’s type are not fully appropriate for the ESL chatbot since the chatbot should converse satisfactorily with an ESL student “out of the box.” However, the application of clouding in accumulating accents deserves exploration. Here are two ways to take advantage of clouding in ESL chatbot design:

1. Use local clouding systems to aggregate different accents of a specific area. Build a sound corpora based on the other three suggestions listed in this section. Integrate the corpora with an ESL chatbot into an offline program. Two advantages emerge from this design: No networking is needed and the accessing speed is admittedly stable.

2. Integrate the above corpora into an online program. A local clouding system may run alongside an online chatting program, retrieving accents unfound before but without intervening in ongoing conversations between the user and the chatbot. One advantage is that the existing corpora, updated intermittently off line, can improve. One downside is that networking has to be readily available.
III. Suggestions for the Improvement of Semantic Recognition Rate

Our suggestions in the previous section concern the construction of soundbases. In this section, textbases, i.e., the chatbot’s built-in replies, come to the forefront. We suggest that the textbases should be level-specific, topic-specific, and scenario-based.

In the level-specific chatbot, a list of topics for conversation is available for the student to select from. These topics are derived from the student’s textbook, closely relating to young beginners’ interests and concerns. After acquiring some basic idea of the topics from the textbook, the student is prepared to engage in conversation with the chatbot through a topic of interest to him/her. For the convenience of discussion, let’s say that the topic presently chosen by the student is “Knowing Animals,” or to be more narrow in scope, “Learning the Names of Animals.” In compliance with the student’s choice, the chatting program offers him/her a scenario, presumably titled “Visiting the Zoo.” In this scenario, the chatbot, or to be exact, the talking avatar, can opt for the role of a zoo guide. She is able to answer questions regarding properties of an animal introduced in the textbook.

Two points are noteworthy in the design of the scenario:

First, the zoo guide, apart from supplying background knowledge of the zoo, should take the initiative in asking questions related to animal properties and resources offered in the textbook. Unable to speak much English, the ESL student normally has difficulty in uttering a question. Accordingly, it is advisable for the chatbot to initiate a conversation. The student may be encouraged to take the leading role later, depending on his/her familiarity with the questioning patterns.

Second, a soft-tone approach is more preferable than a hard-tone one. Hard tone here means using the principle of positivity (yes/right) and negativity (no/wrong) in assessing a response from the user. What matters in an ESL conversation, in our opinion, does not reside in the student’s answer being right or wrong but his/her continuing engagement in the conversation. Contrary to the hard-tone approach, a soft-tone one uses responses of “conditional positive regard” (CPR), a concept modified from Carl Rogers’s “unconditional positive regard” (UPR), which we will address later. CPR refers to an attitude bolstering the student to stay, to talk, and to answer in times when he/she has difficulty finding the right words to say or has nothing to share. As a whole, the avatar is always initiative and active in a conversation and ready to provide assistance, such as cues and clues for further engagement. Consider the following conversation:

The zoo guide: WHAT ANIMAL IS THIS? [The visual avatar is pointing at a
zebra.]
The ESL student: This is a horse. [The student does not know how to say the word, “zebra,” or has difficulty saying it.]
The zoo guide: THIS IS AN ANIMAL LIKE A HORSE. TRY ONE OF THE FOLLOWING:
A. THIS IS A LION.
B. THIS IS A ZEBRA.
C. THIS IS AN ELEPHANT.

[Possible Dialogue #1:]
The ESL student: a. This is a lion.
The zoo guide: YOU ARE CLOSE. TRY AGAIN.
The ESL student: b. This is a zebra.
The zoo guide: YOU’RE RIGHT. LET’S MOVE ON . . . .

[Possible Dialogue #2:]
The ESL student: b. This is a zebra.
The zoo guide: YOU’RE RIGHT. LET’S SAY IT AGAIN. THIS IS A ZEBRA.
The ESL student: This is a zebra.
The zoo guide: VERY WELL. LET’S MOVE ON . . . .

In response to the wrong answer, “This is a horse,” the zoo guide does not strike out the student with a NO. Instead, she walks the student forward with an analogical clue. Next, she provides options for the student to choose from, instead of letting him/her run idle, with little idea where to go next. These options constitute a cue, something that excites or stimulates the student to continue.

To sum up, questions initiated by the chatbot serve to guide the student in learning, leading him/her toward a satisfactory level of proficiency before he/she assumes an active role in the conversation. A response that carries clues and cues, i.e., a soft-tone approach, helps motivate the student to move on.

IV. What Goes Awry in Siri’s Processing a Weird Statement

In darkesseofmagyk’s movie clip, when the user asks, “I need to hide a body,” Siri displays 5 locations for the mission: 1. mines; 2. reservoirs; 3. metal foundries; 4. dumps; 5. swamps. Right after the user has picked “swamps” from the list, Siri brings
forth a map, saying, “This swamp looks pretty close to you.” The user’s out-of-context request is a purely humorous move, given its weirdness. Unable to detect the humor embedded in the interrogator’s utterance, Siri takes it at face value, like a request for help in finding a restaurant in the neighborhood. What goes awry in Siri’s processing this input? How does an ESL chatbot designer tackle weird statements like “I need to hide a body”? We will address the first question in this section and the second in the next. In this section, after clarifying key factors of Siri’s silly response to that weird “context-independent” statement, we will bring in “context-dependency,” a topic related to the issue of how to improve an ESL chatbot’s accuracy rate in responding to simple but meaningful statements.

The input-output process of an STS chatbot, partly derived from Stanford, Williamson, Sherwin, & Castellucci (1997), can be summarized as follows:

speech input ⇔ speech recognition ⇔ speech-to-text transcribing
⇔ semantic recognition (choosing a textual response from a textbase)
⇔ text-to-speech transcribing ⇔ speech output

This simplified flow chart quickly shows that Siri’s problem originates from the semantic recognition section. darkesseofmagyk’s inquiry, along with Wozniak’s (“What are the 5 biggest lakes in California?”), mounts no challenges to Siri in speech recognition, but they have plunged Siri into a semantic jungle. A comparison indicates that, in Wozniak’s case, Siri has too many options to choose from the textbase in spite of the fact that Wozniak’s question is a simple and meaningful one, and, in darkesseofmagyk’s case, Siri has no options to choose from since “I need to hide a body” is semantically ambiguous and routinely unusual. For further discussion, an explanation about a chatbot’s mechanism of selecting a response from its textbase is in order. Mostly, a chatbot uses pattern matching and pattern recognition algorithms to discern the meaning of an input. According to the article, “Pattern Recognition,” pattern matching algorithms “look for exact matches in the input with pre-existing patterns. A common example of a pattern-matching algorithm is regular expression matching, which looks for patterns of a given sort in textual data.” When an input meets no pre-existing patterns to match, pattern recognition algorithms weigh in. They will calculate all the keywords or parts-of-speech available in the input and produce a pattern most approximate to the input and subsequently gives a response based on the chosen pattern. This selecting method can be called “fuzzy matching,” in contrast to “exact (pattern) matching.” It is not possible that a chatbot’s textbase has all the patterns available for retrieving. Therefore, it is not surprising that Siri finds no patterns to match a sentence like “I need to hide a body.” An
approximate pattern, instead, is evoked to serve the conversational purpose. For illustration, let’s say the approximate pattern is “I need to hide something.” Hiding, inferentially, involves locations and subsequently Siri churns out a list of places.

What does “a body” refer to in “I need to hide a body”? A human corpse? A tiger’s carcass? This semantic ambiguity, coupled with Siri’s fuzzy matching algorithms, has thrown Siri into an awkward predicament. “I need to hide a body,” a context-independent sentence, is unclear in meaning. By contrast, “What is your favorite color?” also a context-independent sentence, is transparent in meaning. Accordingly, a pattern for the latter can be easily identified and a suitable answer to the question pops out in no time. But reality is more complex than this. Weird questions excluded, simple and meaningful questions are sometimes difficult to deal with, not because of the depth in lexical and compositional meaning but because of context-dependency. Consider the first three questions provided by Jurafsky and Martin (2009) in their introduction of “Web-based question answering”:

1. What does “divergent” mean?
2. What year was Abraham Lincoln born?
3. How many states were in the United States that year?

Question 1 is a “definition question” while Question 2 is a “factoid question.” It is easy to answer them, according to Jurafsky and Martin (2009, p. 36). Note that these two questions are context-independent. Question 3, a meaningful sentence, however, can pose lots of problems for recognition algorithms to find a pattern because it is context-dependent. The first part of the question, “How many states were in the United States,” seems a factoid question. But, “What year is that year?” Jurafsky and Martin (2009, p. 37) ask. They continue: “To interpret words like that year, a question-answering system needs to examine the earlier questions that were asked; in this case, the previous question talked about the year that Lincoln was born.” It should be noted that parsing previous questions for an answer, a method called “coreference resolution,” may not work satisfactorily when a conversation involves more than one topic. In a multi-topic context, many things could have happened and existed in that year. It is accordingly very hard for a textbase to contain all the information about them. To make coreference resolution an effective method, the conversation should be topic-specific, along with those specificities suggested earlier. Drawing from the discussion, we suggest that:

1. Context-independent but semantically meaningful questions for an ESL student such as “What is your favorite animal?” are preferable in the initial
stages of learning. This design fosters an ESL student’s confidence. Note that context-independent questions may wear out an ESL student quickly.

2. Context-dependent questions should enter into the scene only after the ESL student has garnered a full acquaintance with the context. To be exact, these questions should be context-specific since they are derived from a single topic-specific scenario. As such, it is easier for the designer, with a small but specific context at his disposal, to map out all the possible patterns.

V. The UPR Approach: Tackling Weird Input from the User

In this section, we advise that a Rogerian approach be used to tackle weird questions or statements from the user. Rogerian psychotherapy is characterized by the therapist’s “unconditional positive regard,” among other qualities. UPR, as defined in an article titled “Person-Centered Therapy,” means that “the therapist accepts the client totally for who he or she is without evaluating or censoring, and without disapproving of particular feelings, actions, or characteristics.” Additionally, as expounded in the same article, UPR is intended for the construction of “a nonthreatening context in which the client feels free to explore and share painful, hostile, defensive, or abnormal feelings without worrying about personal rejection by the therapist.” This UPR attitude, transfigured for ESL conversation practice, bears a resemblance to the CPR attitude at first glance since both carry a soft tone. To be concise, as is elaborated below, the UPR approach is meant to address a different issue than the one handled by the CPR approach.

Typically, when a chatbot finds no patterns in its textbase that match user input, it randomly grabs one from a particular repository, called the “Ultimate Default Category” by Wallace (2005), the maker of ALICE, an online general-purpose chatbot. Such random responses sometimes capture a bit of humor but mostly disrupt a conversation as non sequiturs. Applied to an ESL chatbot, random responses might cause the ESL student to quit or get lost in a conversation. As an alternative to randomness, we recommend that a Rogerian mode of inquiry be adopted to augment the chance of the user’s continuing to engage in the dialogue—specifically, the Rogerian inquiry method of quoting a patient’s dictum in the interrogation of a patient. This approach was adopted by Weizenbaum’s “Eliza” (1966), one of the earliest TTT chatbots. Below is a sample of a dialogue with Eliza (Weizenbaum, 1966, p. 37; quoted in Wardrip-Fruin, 2009, p. 25-26; words in upper case are Eliza’s responses):

Men are all alike.
IN WHAT WAY
They’re always bugging us about something or other.
CAN YOU THINK OF A SPECIFIC EXAMPLE
Well, my boyfriend made me come here.
YOUR BOYFRIEND MADE YOU COME HERE
He says I’m depressed much of the time.
I AM SORRY TO HEAR YOU ARE DEPRESSED
It’s true. I am unhappy.
DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE UNHAPPY
I need some help, that much seems certain.
WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP
. . . .
You are not very aggressive but I think you don’t want me to notice that.
WHAT MAKES YOU THINK I AM NOT VERY AGGRESSIVE
You don’t argue with me.
WHY DO YOU THINK I DON’T ARGUE WITH YOU
You are afraid of me.
DOES IT PLEASE YOU TO BELIEVE I AM AFRAID OF YOU
. . . .

Eliza’s responses above illustrate a full embodiment of UPR. Her statements divulge little vein of positivity or negativity. A close look reveals that the user’s utterances are borrowed and repeated in Eliza’s responses in a slightly altered manner, heralded by neutral interrogative phrases such as “WHAT WOULD IT MEAN TO YOU IF . . . ,” “WHAT MAKES YOU THINK . . . ,” “DOES IT PLEASE YOU TO BELIEVE . . . .” According to Wardrip-Fruin (2009, p. 30), “Each statement by an Eliza script is the result of a multistep transformation of the most recent audience statement.” Noteworthy is that in the multistep transformation, Eliza has never expressed opinions of her own. What counts eventually is that she succeeds in making the user willing to continue articulating himself/herself.

To computer scientists, the internals of the Eliza software are not very complex, but Eliza’s success is so enormous that most users assume it runs a sophisticated system like HAL, a sci-fi AI recounted in Clarke’s 2001: A Space Odyssey (1968) and made famous by the movie of the same title. Some people call this assumption of complex software “The Eliza Effect” (Wardrip-Fruin, 2009, p. 31-32). This term implies a shortage of AI in Eliza, but, ironically, it is what an ESL chatbot needs since it hooks an ESL student and motivates him/her to continue the practice.

An imaginary dialogue of our ESL chatbot in the vein of Eliza could look like the
We presume that after the ESL student has gained more linguistic knowledge of the animal topic and the related scenario, he/she might sometimes ask questions or make statements beyond the range of the chatbot’s established patterns.

I want to kill lions.
TELL ME MORE ABOUT IT
They eat meat.
WHAT ELSE COMES TO MIND WHEN YOU THINK OF THEM
They are bad.
WHAT MAKES YOU THINK THEY ARE BAD
Horses and zebra is afraid of them.
WHY DO YOU THINK HORSES AND ZEBRAS ARE AFRAID OF THEM
Lions has big teeth.
ARE YOU AFRAID OF BIG TEETH

Note that we presume that a disambiguation device is in place in the imaginary chatbot to fix those grammatically incorrect words or phrases. That said, this dialogue sample evidently adopts the UPR approach. “I want to kill lions,” an emotional statement, presumably, will not be likely to have a pre-existing pattern to match in the textbase because of its spontaneity or unexpectedness, which are typical of real-life conversations. Instead of feeding the ESL student with a random response, risking rupture of the conversation, the chatbot serves up a UPR statement, which has a high likelihood to sustain the learner’s engagement. Subsequently, the chatbot keeps “recycling” the user’s statements. This method is called “reflection” in “Person-Centered Therapy”: “This technique shows that the therapist is listening carefully and accurately, and gives clients an added opportunity to examine their own thoughts and feelings as they hear them repeated by another person. Generally, clients respond by elaborating further on the thoughts they have just expressed.” In our case, this method is modified into a technique to induce the ESL student to keep on talking. The key feature of this technique is feeding the ESL student with statements whose syntactic format mostly comprises two parts: neutral interrogative phrases + recycled words/phrases from the user.

The UPR approach and the CPR approach do not differ much from each other. What mostly distinguishes them are:

1. A UPR chatbot, like Eliza, does not express opinions, positive or negative,
while a CPR chatbot offers opinions, including clues and cues, and its attitude involves positivity but not negativity.

2. The CPR approach is used to tackle situations wherein an ESL student is experiencing difficulty in expressing himself/herself. The UPR approach is used to handle an ESL student’s weird input, such as “I need to hide a body,” “I want to kill lions,” or even “Colorless green ideas sleep furiously,” a “grammatically correct but semantically nonsensical” sentence made by Chomsky (1957, p. 15).

Note that if the student’s input is constantly “grammatically incorrect and semantically nonsensical,” like Chomsky’s “Furiously sleep ideas green colorless,” the chatbot may demand the student to review his/her textbook before another try. It is also noteworthy that the ESL chatbot may shift between the CPR and UPR approaches in the engagement, depending on the student’s input or responses.

VI. Conclusion: Three Steps towards a Better ESL Chatbot

Since the imaginary ESL chatbot, based on a particular textbook, is content-specific, pattern matching algorithms can easily assume most of the responsibilities in identifying user input with exact patterns. Amid such a conversation, the CPR approach is resorted to only when the ESL student experiences difficulty in finding the right word to say. In the case of illogical or weird input, fuzzy matching algorithms step in first to figure out an appropriate pattern. If fuzzy matching fails to submit one, it is advisable to apply the UPR approach, rather than the Ultimate Default Category. Amid a UPR conversation, if a statement used by the student matches a pattern in the textbase, the previous normal processing can then resume. Theoretically, this three-step process, consisting of pattern matching, fuzzy matching, and the soft-tone (CPR/UPR) approach, guarantees a higher success rate for an ESL chatbot to prolong a student’s engagement in conversation practice.

References


