SECOND LANGUAGE ACQUISITION THROUGH HUMAN COMPUTER DIALOGUE

Stephanie Seneff, Chao Wang, Mitch Peabody, and Victor Zue

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA
{seneff,wang,mizhi,zue}@csail.mit.edu

ABSTRACT

This paper describes our recent research in developing tools for second language acquisition based on spoken dialogue interaction with a computer. We argue that language proficiency can best be achieved through active communication, and that the computer is very patient and provides a non-threatening environment in which to practice. We have adapted our pre-existing multilingual dialogue systems for this application, focusing in our initial prototype on an English-speaking student learning Mandarin within the weather domain. Two significant new contributions are a Web-based interface for practice exercises to gain proficiency in carrying out a live conversation and a high-quality narrow-domain spoken language translation capability. In an evaluation on 695 utterances drawn from a corpus of English weather data, our translation system produces an incorrect result less than 2% of the time, with a rejection rate of 8%.

1. INTRODUCTION

Spoken language is the primary means of communication for humans. It is natural (requiring no special training), flexible (freeing hands and eyes to attend to other tasks), and efficient (enabling transmission at rates higher than written language). However, spoken language is as pervasive as it is diverse.[1] Not being able to speak the same language prohibits communication between two people, incurring high cultural, social, and economic costs in our increasingly interconnected world.

Acquiring a second language requires significant effort. To start with, one needs to learn a new set of phonemes, words, and linguistic structures to form sentences and paragraphs. To utilize the target language for communication, one also needs to learn to engage in a conversation, in which discourse, dialogue, and social norms play a crucial role. This is best accomplished with a teacher who knows the language intimately, and who can interact with the student in a variety of settings. However, there are simply not enough teachers to accommodate the numerous people who might be interested in learning a popular second language. Therefore, there is a pressing need for technologies and systems that can help with second language learning.

The use of human language technologies (HLTs) to aid second language learning has been a topic of active research for many years. Despite some successes, some researchers [4, 6, 8] including ourselves, have come to believe that one needs to go beyond the mechanics of standard reading/speaking exercises. Specifically, we need to provide the ability for the system to engage a student in role-playing scenarios (e.g., looking for a hotel with a specific type of accommodations, checking for flight and train schedules, inquiring about weather conditions). This way, students can practice their passively acquired language skills in an active setting, in which they must learn to use the language correctly in order to obtain the desired responses. This environment, we suspect, will enable students to practice interactions in a risk-free setting.

Our work on developing a second language learning system builds on our research on multilingual conversational systems over the past fifteen years [12, 15, 3, 11, 14]. We have created an architecture in which HLT components (speech recognition/synthesis, language understanding/generation, and discourse and dialogue modeling) can communicate with one another through a programmable hub [10]. A key feature of our system is the representation of meaning, which we call a "semantic frame." It serves many roles in our spoken language systems, including database access, discourse maintenance, and language generation.

The remainder of this paper describes our recent work in developing a system for second language learning. We will first describe the overall framework, and give an example of a conversational interaction. This is followed by a section on a Web-based drill exercise and a description of the core translation technology, along with some performance results. We conclude with a brief look to the future.

2. DIALOGUE INTERACTION FRAMEWORK

To incorporate interactive dialogue into a language learning curriculum, we envision a set of lessons related to a general topic (e.g., travel), each focusing on a different scenario (e.g., looking for a hotel, checking on flights, inquiring about weather, and finding a restaurant). Our initial experiments concern interaction in the weather domain, mainly because this is our most mature multilingual system [12, 15, 16]. While our main language pair is Mandarin-English, we also have support for spoken communication in Japanese [15] and Spanish. Since all of our weather information sources are in English, a language translation component renders the weather reports in well-formed Spanish, Japanese, and Mandarin. In fact, the weather reports are preprocessed when harvested into semantic frames and paraphrased back into English even for our English-based system.

This work was supported by the Cambridge MIT Institute, by ITRI, by the Defense Language Institute, and by the National Science Foundation.

[1] According to the Ethnologue [5], there are more than 6,700 languages being spoken by more than 6 billion people in some 230 countries.

Our multilingual systems can be configured to support seamless language switching among the supported languages. In a Mandarin-English configuration, hypotheses for both languages compete in a common search space in the recognizer, and the natural language understanding (NLU) component simultaneously resolves both the meaning of the utterance and the language spoken from a possibly mixed-language N-best list.

To enhance a multilingual system for the language learning application, it is straightforward to reconfigure the control such that an utterance spoken in the native language, L1 (in our case, English) is translated into the foreign language, L2 (in our case, Mandarin Chinese) whereas one spoken in L2 is answered in L2. Thus if, while conversing with the system, the student can not remember how to say something in L2, they can speak an L1 utterance of similar meaning, and then attempt to immediately repeat the translation provided by the system.

A technology component that is newly required for the language learning system is the ability to translate user queries from L1 to L2. The approach we have taken is to use the semantic frame, which encodes the linguistic structure of the user query, as an interlingua [13]. Thus, translation is achieved by generating an L2 string from the semantic frame, using formal generation rules [1].

User:	bo1 shi4 dun4 jin1 tian1 hui4 xia4 yu3 ma5?
	(Will it rain in Boston today?)
System:	tian1 qi4 yu4 bao4 jin1 tian1 bo1 shi4 dun4
	mei2 you3 yu3. (No rain in Boston today)
User:	What is the temperature?
System:	qi4 wen1 shi4 duo1 shao5?
User:	qi4 wen1 shi4 duo1 shao5?
System:	bo1 shi4 dun4 jin1 tian1, zui4 gao1
	qi4 wen1 er4 she4 shi4 du4, jin1 tian1 ye4 jian1,
	zui4 di1 qi4 wen1 ling2 xia4 wu3 she4 shi4 du4.
User:	Could you translate that?
System:	In Boston today, high 2 Celsius,
	tonight, low -5 Celsius.

Fig. 1. Example spoken dialogue between the user and the system. Chinese sentences are represented using tone-marked pinyin.

2.1. Example Conversation

An example of a conversation with the system is shown in Figure 1. At the computer, a user would be able to either type the queries into a type-in window or speak them into a microphone. Because of our carefully designed translation procedure (see Section 4), any translations provided by the system for English inputs are understood by the Mandarin NLU system. We are hoping that this feature will allow even a relatively inexperienced student to converse effectively with the system.

3. WEB-BASED TRAINING EXERCISE

Prior to being able to converse with the system, a student would need to first prepare for a live conversation by studying the vocabulary and sentence constructs appropriate to the lesson's topic. To make this study phase more engaging and more relevant, we have developed some on-line exercises designed to help them master the language usage within the scope of the lesson's topic. Our exercises are centered initially on a native English speaker learning

Fig. 2. Screen dump of Web-based drill exercise, in which the student must solve a weather scenario. The student is provided explicit feedback on any tone errors.

Mandarin [7], and they serve both to provide an active mechanism for the student to absorb the material and to solicit recorded speech and text data to assist the training of our speech recognizer. The interface prompts the user with specific weather scenarios ("Boston – tomorrow – rain") and requires them to type the appropriate query in pinyin format, as illustrated in Figure 2. A follow-up oral exercise requires them to formulate and record spoken queries based on a visual prompt specifying the same sequence of scenarios.

One of the most difficult aspects of Mandarin for an English speaker is the use of prosodics to encode the five tones, which contribute in important ways to the lexical content (for example, the syllable "tang" could mean "soup," "sugar," "lie down," or "hot," depending only upon differences in tone.) It is unrealistic to expect the students to have perfect tone knowledge. Robustness in understanding sentences with tone errors is achieved by building a word graph out of the student's typed inputs, substituting, for each syllable, alternates that cover all supported variants with respect to tone. The NLU system is already capable of searching a graph for the highest scoring hypothesis, which is in fact its standard mode for processing recognition word graphs. The NLU system verifies both the syntactic and semantic correctness of the input by parsing the input and comparing the extracted key-value information against the prompt.

We have decided to configure our two type-in exercises such that the first one concentrates on phonetic knowledge and the second one emphasizes tonal knowledge. In both exercises, the student is *not* required to mark the tones correctly in order to be understood – otherwise the exercise would be far too frustrating. However, in stage 2 tone errors are completely ignored, whereas in stage 4 the student's sentence is presented alongside a repaired version where syllables with incorrect tones are highlighted in red, as illustrated in the screen dump in Figure 2.

As schematized in Table 1, our 5-stage drill exercise alternates

	baseline	no tone		with tone	
	Stage 1	**Stage 2**	**Stage 3**	**Stage 4**	**Stage 5**
Prompts	pinyin text string	scenario (e.g. "Boston – rain – tomorrow")			
Mode	speech	text	speech	text	speech
Tone	assessed off-line	ignored	assessed off-line	corrected	assessed off-line

Table 1. 5-stage drill exercise designed both to prepare the student to converse with the system and to solicit text and audio data.

between speech mode and text mode, and the task becomes progressively more difficult. We are assuming that the student has some knowledge of Mandarin initially, and is able to read and write in pinyin. Stages 1, 3, and 5 involve recording spoken utterances. In stage 1, the student is simply asked to read aloud a set of questions on topics related to weather, represented in tone-marked pinyin (e.g., "qi4 wen1" for "temperature.") This initial stage will establish a benchmark measure of the student's understanding of the mapping from the symbolic tone representation to its spoken encoding. In stages 3 and 5, they must compose the weather questions from provided English prompts, such as "Shanghai – windy – Friday." Stages 2 and 4 involve typing into a Web page, and are intended to prepare them for the subsequent audio recording stages. These type-in stages are also based on composing sentences from prompts providing different weather scenarios (randomized to new scenarios each time the exercise is repeated). However, the student has the opportunity to type any weather-related phrase or sentence in English in order to obtain a translation into Mandarin. The prompts for the audio stages 3 and 5 are identical to those in stages 2 and 4, respectively.

We expect that the recordings obtained from stages 1, 3, and 5 will provide data that will be valuable not only as training data for the recognizer and NLU components, but also for the purpose of assessing the effectiveness of our second text-based drill exercise (stage 4) in teaching tone. The initial recordings will inform us of how well the student understands the mapping from the symbolic tone representation to its prosodic encoding, a necessary precondition to improvement through explicit lexical knowledge. The hope is that the student will retain knowledge of lexical tone from explicit corrections of their errors, which will be reflected in improved tone accuracy in their subsequent recordings.

4. SPEECH TRANSLATION

Language learning presents special challenges to a translation system because the quality of the translation must be essentially perfect, to avoid teaching the student inappropriate language patterns. However, it is only essential that the translation be fluent and contain the same overall meaning as the original L1 utterance. It need not be an exact translation, but could instead be viewed more as a paraphrase. In fact, we believe that it would be beneficial to the student if there was some randomness in the translation, such that multiple repetitions of the same English utterance produced different ways to render an equivalent intent in Mandarin. Interestingly, since the student is very likely to repeat verbatim what the system proposes, it is of paramount importance that the translation be understandable by the system's L2 grammar.

Rule-based Translation: To generate well-formed strings in L2, we utilize the GENESIS language generation framework [1]. It works from a lexicon providing context-dependent word-sense surface strings for each vocabulary item, along with a set of recursive rules specifying the ordering of constituents. Variability in the surface form can be achieved by randomly selecting among alternative rules and lexical entries. For example, an English query such as "Will it snow in Chicago this weekend?" can be realized either as the Chinese "statement + question-particle" (... hui4 ... ma5) construct or as the "A-not-A" (hui4 bu2 hui4) construct, with additional permutation on the ordering of the time and location phrases. This is useful not only to the language student, but also to the system, since we can generate a rich set of Chinese sentences for training the language models of the speech recognizer.

```
{c verify
   :auxil "will"
   :topic {q pronoun
             :name "it" }
   :pred {p rain
          :pred {p locative
                 :prep "in"
                 :topic {q city
                           :name "boston" }}
          :pred {p temporal
                 :topic {q weekday
                           :quantifier "this"
                           :name "weekend" }}}}
```

City: "Boston" **Date:** "weekend" **Topic:** "rain"

Fig. 3. Semantic frame and derived E-form for the example sentence "*Will it rain in Boston this weekend?*"

The generation rules can be fine-tuned by experts to produce high-quality outputs on a set of development data. However, when the input deviates from the expected patterns, either due to novel linguistic constructs or caused by speech recognition errors, the rule-based generation module could produce ill-formed outputs. Thus, to assure quality control and to maximize coverage, we have configured a system which uses verification of parsability by the L2 grammar to accept/reject a generated L2 string. If the rule-based procedure fails to generate a parsable query, we invoke an example-based back-off mechanism to select an utterance with equivalent overall meaning, but which may deviate considerably from the original with regard to syntactic structure.

Example-based Translation: We have available to us a large corpus of English weather queries, recorded from phone calls to the publicly available Jupiter weather system [16]. These can be used to create a table of examples for the example-based framework. To index the examples, we can conveniently exploit a pre-existing capability to translate from the original hierarchical semantic frame into a flattened "E-form" ("electronic form") structure containing a succinct set of [key: value] (KV) pairs associated with the utterance. An example semantic frame and derived E-form for the utterance "will it rain in Boston this weekend" is shown in Figure 3.

The E-form can be further generalized by replacing specific values with class tags for selected attributes, for instance, substitut-

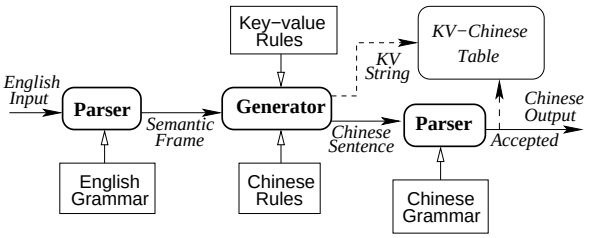

Fig. 4. Schematic diagram of process to both verify quality of rule-based generation and create a table for the example-based back-off framework. Note: KV = [Key: Value].

	Perfect	Adequate	Wrong	Failed	%
Rule	550	34	8		85%
Example	27	16	5	55	15%
Total	577(83%)	50(7%)	13(2%)	8%	100%

Table 2. Performance evaluation of the rule-based and example-based translations for 695 utterances selected from our weather domain corpus.

ing <city_name> for "Beijing" and "<weekday>" for "Tuesday." A corpus of English utterances can be pre-selected for parsability in L2, and then indexed with the associated class-reduced E-form.

Figure 4 summarizes the process which both confirms the quality of rule-based translations and prepares a database to serve as a translation memory for the example-based method. In using this database for translation, an example utterance with a matching E-form is adjusted by replacing class-assigned attributes with the value instantiated from the original utterance. The database of examples can also be augmented with any available original Chinese data, in which case the KV index can be derived using the Chinese grammar for parsing.

5. EVALUATION

To evaluate our translation framework, we used a set of 695 utterances, selected from held-out telephone recordings. Utterances whose manually-derived transcription can not be parsed by the English grammar are excluded from the evaluation, since they are likely to be out-of-domain sentences and would simply contribute to null outputs. The SUMMIT speech recognizer [2] and the TINA NLU system [9] were used to produce the semantic frame encoding the utterances' linguistic content. The recognizer achieved 6.9% word error rate and 19.0% sentence error rate on this set. The test data have on average 6.5 words per utterance.

Table 2 summarizes our results from this experiment. A bilingual judge rated the translation quality as "perfect," "adequate," or "wrong," based on both grammaticality and fidelity. The rule-based system was preferred if it produced a sentence that was parsable by the Mandarin grammar, accounting for 85% of the data. The example-based system was invoked for the remaining 15%, but it failed to find a match for over half of these. However, when it did find a match, it obtained a reasonable translation nearly 90% of the time, increasing the overall yield by 6%. Most significantly, an incorrect translation occurred only 2% of the time.

6. SUMMARY AND FUTURE PLANS

This paper has described our recent research on the idea of exploiting multilingual dialogue systems for language learning applications. We have developed a Web-based interface to help the student to prepare for conversational interaction. In addition, we have developed a high-quality spoken language translation capability to assist the student interactively during the dialogue. Our immediate future plans are to launch a data collection effort using these tools, working with students at the Defense Language Institute in Monterey, California. Long term plans involve extensions to other domains and languages, including the development of tools to empower language teachers to design lesson plans based on a wide range of possible topics.

7. REFERENCES

[1] Baptist, L. and Seneff, S. "Genesis-II: A Versatile System for Language Generation in Conversational System Applications," *Proc. ICSLP '00*, III, pp. 271–274, Oct. 2000.

[2] Glass, J., Chang, J, and McCandless, M., "A Probabilistic Framework for Feature-based Speech Recognition," *Proc. ICSLP*, IV, pp 1–4, 1996.

[3] Chuu, C., "Lieshou: A Mandarin Conversational Task Agent for the Galaxy-II Architecture," MIT MEng Thesis, Dec. 2002.

[4] Eskenazi, M., "Using Automatic Speech Processing for Foreign Language Pronunciation Tutoring: Some Issues and a Prototype," Language Learning and Technolgy, 2 [2], pp. 62-76, Jan. 1999.

[5] Grimes, B. (ed.), *Ethnologue*, Summer Institute of Linguistics, Academic Publications, 1992.

[6] Holland, V. M., Kaplan, J. D., and Sabol, M. A., "Preliminary Tests of Language Learning in a Speech-Interactive Graphics Microworld," Calico Journal, 16 [3] pp. 339-359, 1998.

[7] Peabody, M., Seneff, S., and Wang, C., "Mandarin Tone Acquisition through Typed Dialogues," pp. 173–176, *InSTIL Symposium on Computer Assisted Language Learning*, 2004.

[8] Raux, A. and Eskenazi, M., "Using Task-Oriented Spoken Dialogue Systems for Language Learning: Potential, Practical Applications and Challenges," pp. 147–150, *InSTIL Symposium on Computer Assisted Language Learning*, 2004.

[9] Seneff, S., "TINA: A Natural Language System for Spoken Language Applications," *Computational Linguistics*, 18 [1], pp. 61–86, 1992.

[10] Seneff, S. Hurley, E., Lau, R. Pao, C. Schmid, P. and Zue, V., "Galaxy-II: A Reference Architecture for Conversational System Development," *ICSLP '98*, pp. 931–934, Dec. 1998.

[11] Seneff, S., Wang, C., and Zhang, J. "Spoken Conversational Interaction for Language Learning," pp. 151–154, *InSTIL Symposium on Computer Assisted Language Learning*, 2004.

[12] Wang, C., Cyphers, D. S,, Mou, X., Polifroni, J., Seneff, S., Yi, J., and Zue, V., "MUXING: A Telephone-access Mandarin Conversational System," *Proc. ICSLP '00*, II, pp. 715–718, Oct. 2000.

[13] Wang, C. and Seneff, S. "High-quality Speech Translation for Language Learning," pp. 99–102, *InSTIL Symposium on Computer Assisted Language Learning*, 2004.

[14] Zue, V., Seneff, S., Polifroni, J. Meng, H., and Glass, J., "Multilingual Human-Computer Interactions: From Information Access To Language Learning," *Proc. ICSLP*, 1996.

[15] Zue, V., Seneff, S., Polifroni, J., Nakano, M., Minami, Y., Hazen, T.J., and Glass, J. "From Jupiter to Mokusei: Multilingual Conversational Systems in the Weather Domain," *Proc. MSC2000*, pp. 1–6, Oct. 2000.

[16] Zue, V., Seneff, S. Glass, J., Polifroni, J. Pao, C., Hazen, T. J., and Hetherington, L. "JUPITER: A Telephone-Based Conversational Interface for Weather Information," *IEEE Trans. on Speech and Audio Proc.* 8 [1], 2000.