A Spoken Translation Game for Second Language Learning [1]

Chao WANG [2] and Stephanie SENEFF
MIT Computer Science and Artificial Intelligence Laboratory

Abstract. In this paper, we describe a Web-based spoken translation game aimed at providing language learners with an easily accessible and fun environment to practice speaking the foreign language. Our prototype centers on the task of translating flight domain sentences from English to Chinese. The system presents English sentences as stimuli to elicit Chinese utterances from a user. It tracks a user's performance and rewards them with advancements in difficulty level. A user study was conducted involving 12 learners of Mandarin Chinese. Each participant played the game and answered survey questions. The system achieved 9.7% word error rate on 2834 non-native utterances collected in the user study. All subjects thought the system was helpful at improving their Chinese, and most of them would play it again and recommend it to their Chinese-learning friends.

Keywords. Computer aided language learning, human language technology, machine translation, educational games, human computer interaction

1. Introduction

How to gain and retain skills in a foreign language is a problem facing many language learners. It is generally agreed that living in a country where the new language is spoken is the best way to become fluent [1]. Unfortunately, a total immersion experience is hard to attain, and most students still learn a foreign language in a classroom setting. [2] provides an excellent summary of conditions and pedagogical recommendations for success in a new language, and eloquently argues for the use of computers in foreign language teaching to overcome limitations of the traditional classroom setting.

Speech and language technologies have great potential in Computer Aided Language Learning (CALL) applications. Ideally, a voice-interactive system can role play a language teacher and a conversational partner, to provide the learner with endless opportunities for practice and feedback. In reality, voice-interactive CALL applications are severely constrained by technology limitations [3]: recognition and understanding of non-native speech and robust dialogue modeling remain as unsolved research challenges. A relatively successful application of speech processing technology is in the area of pronunciation training [4,5,6]. In this case, a learner repeats words or phrases prompted by the computer, and receives feedback on the quality of their phonetic pronunciation and

[1] This research is supported by ITRI in Taiwan and by the Cambridge MIT Institute.

[2] Corresponding Author: Chao Wang, 32 Vassar Street, Room 362, Cambridge, Massachusetts 02139, USA. Tel.: +1 617 253 7772; Fax: +1 617 258 8642; E-mail: wangc@csail.mit.edu.

intonation. In contrast, dialogue systems allow language learners more freedom in constructing their own sentences to practice conversational skills. While a number of dialogue systems have been developed (or adapted) for language learning purposes [7,8], the performances are typically too brittle to be widely used by end-users.

This paper explores the use of speech translation technology to develop a novel voice-interactive CALL application [3]. We describe a Web-based computer game, which aims to provide an inviting environment for language learners to practice speaking the new language. By using Web technology [9], the game is easily accessible to anyone with a computer equipped with a microphone and Internet access. Furthermore, no specialty software is required besides a Web browser and the Java run time environment.

Our game prototype is centered around the task of translating spoken phrases and sentences between English and Chinese in the flight information domain. We envision that the translation exercise would serve as a preparation phase for later dialogue interaction with a conversational system for booking flights [10]. The game uses English sentences as stimuli to elicit Chinese speech from the learner. The system keeps track of how many turns a learner takes to complete all the sentences in a game session, and rewards good performance by advancing the learner towards higher difficulty levels. Through the use of speech recognition, language understanding and language generation technologies, the system is able to provide immediate feedback to the learner by paraphrasing his or her utterances in both languages and judging if the perceived translation is correct [11]. The learner can also get help from the system, by asking for a Chinese translation of an English phrase or sentence. The game system utilizes an interlingua-based bidirectional translation capability, described in detail in [12,13].

In the following sections, we will first describe the game system, including its basic features and the underlying technology components. Next, we will report findings of a user study involving 12 participants, who interacted with the system and filled out a survey. We conclude with future plans to extend our work.

2. System Overview

When a student accesses the game's URL, a login page is first presented, where they can enter a user name (for tracking performance), as well as optionally specify the difficulty level and the number of sentences to work on for each game session. Subsequently, the main page (shown in Figure 1) presents a randomly generated task list, and the system prompts the user to translate each sentence in turn. The system paraphrases each user utterance in both languages to implicitly inform the user of the system's internal understanding, and judges whether the student has succeeded in the task. If the user correctly translates the sentence, the system congratulates him/her and advances to the next sentence. Otherwise, the student can click the meta command buttons (e.g., "help me" or "give up") or simply try again. Once the student completes the sentence list, the system summarizes their performance and offers the option to play another round, possibly at a different level of difficulty (up or down). Figure 2 shows an example interaction between the user and the system.

[3] We are not aware of other CALL applications based on a similar paradigm.

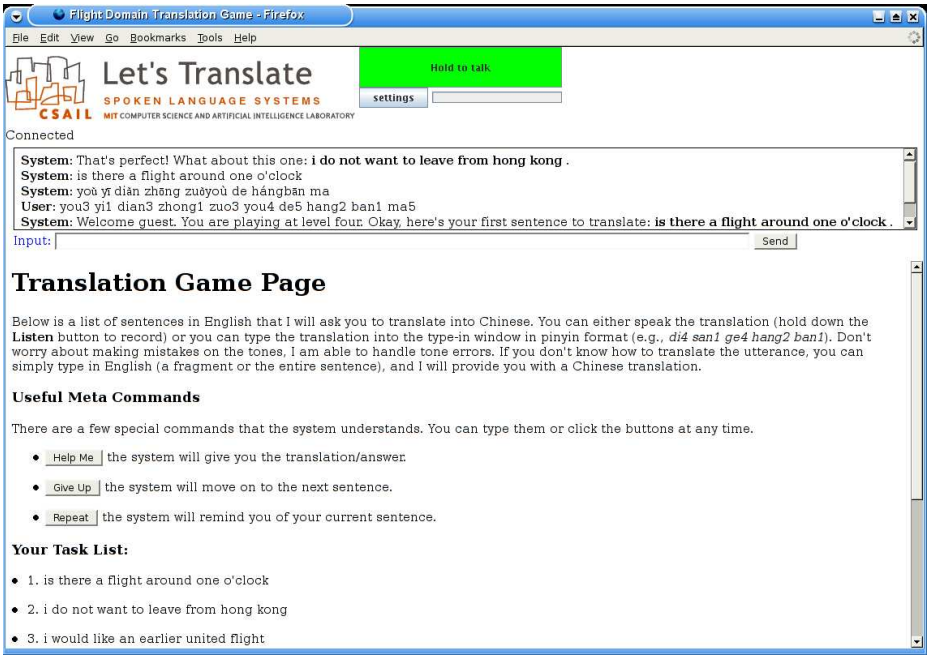

Figure 1. Web interface for the translation game.

System:	(*greet and prompt user*) Welcome! You are playing at level four. Okay, here's your first sentence to translate: **The return date is Tuesday October 25**.
User:	(*translate with error*) fan3 hui2 ri4 qi1 shi4 si4 yue4 er4 shi2 wu3 hao4 xing1 qi1 er4.
System:	(*paraphrase in Chinese*) fan3 hui2 ri4 qi1 si4 yue4 er4 shi2 wu3 hao4 xing1 qi1 er4.
System:	(*paraphrase in English*) The return date is Tuesday April 25.
User:	(*repeat date*) shi2 yue4 er4 shi2 wu3 hao4.
System:	(*paraphrase in Chinese*) shi2 yue4 er4 shi2 wu3 hao4.
System:	(*paraphrase in English*) October 25.
System:	(*respond to user*) "Great! You successfully translated the sentence in separate fragments. Let's move on to the next sentence: **I need to fly to Hong Kong the day after tomorrow**.
User:	(*click on the help button*)
System:	(*repeat English prompt*) I need to fly to Hong Kong the day after tomorrow.
System:	(*offer example translation*) wo3 xu1 yao4 hou4 tian1 fei1 xiang1 gang3.
User:	(*imitate the translation*) wo3 xu1 yao4 hou4 tian1 fei1 xiang1 gang3.
...	...
System:	(*respond to user*) You translated nine out of ten sentences. You took on average 1.6 turns per sentence. You have advanced to level five. Would you like to play another round?

Figure 2. Example interaction between a user and the system.

2.1. User Utterance Processing

A user's utterance is first processed by the speech recognizer, as shown in Figure 3. The recognizer is based on the SUMMIT landmark-based speech recognition system [14], using Chinese acoustic models trained on native speakers' data [15,16]. Mandarin tone features are ignored in the acoustic models; however, the language model implicitly captures some tone constraints. This is perhaps more appropriate than modeling tone explicitly

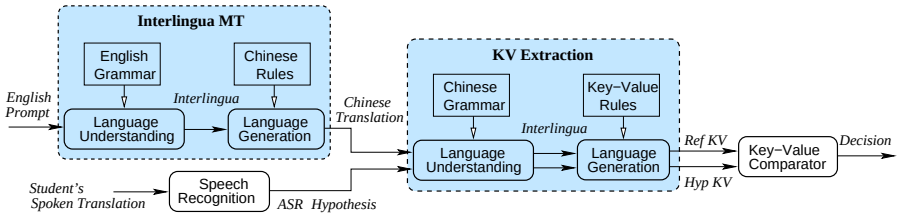

Figure 3. Flow chart of the user utterance verification process.

in the acoustic model, considering that non-native speakers typically have trouble pronouncing tones. The language model was initially trained on Chinese translations of English sentences used in the game. While this achieves good performance when the learners use sentence patterns expected by the system, the performance degrades significantly when they deviate from the expected patterns. We devised a semi-automatic procedure to improve the language model data. The Chinese sentences were converted into templates by using the parser [17] to replace selected words and phrases with variables. This resulted in a much more compact representation of the sentence patterns. We then manually examined the Chinese templates, augmenting the patterns with appropriate alternatives. This process can be iterated as real user data are collected.

After a user's Mandarin utterance is recognized, it is evaluated to determine whether it conveys the same meaning as the English stimulus. This is done via a series of steps as illustrated in Figure 3. The English prompt is first translated into Chinese, and this translation is then compared with the user's translation via an encoding of the meaning in terms of [key: value] (KV) pairs, for ease of scoring. Mismatches are tabulated into substitutions, deletions, and insertions. A perfect match is achieved if the KV pairs of the user sentence and reference are identical. However, partial matches are common, especially when substitution errors on dates and times occur as a result of misrecognition. In such situations, it is natural for the user to just repeat the "incorrect" piece, especially when the sentence is long. Hence, a partial match mode was added to the comparison algorithm, to allow the student to complete the translation in multiple turns. Further details for the assessment algorithm and evaluation results are described in [11].

When the recognizer makes errors, it could result in the system rejecting a perfectly fine user sentence or accepting a wrong answer. Although both types of errors are unfavorable, the first type happens more often and is potentially more frustrating to the user. A well-formed user sentence could even fail in repeated attempts, because of inadequacies in recognition and/or understanding. We provided two mechanisms to work around this condition: the system will move on to the next stimulus sentence if the user is stuck on a sentence for more than a certain number of turns, and the student can also use a meta command to "give up" on the intractable stimulus.

2.2. Game Sentence Inventory

The game system uses a total of over 1000 templates of English sentences in the flight domain, from which it generates the "task list" for each game session. The templates were organized by their length, from short to long, generally reflecting increased difficulty levels. Some manual effort has been devoted to adjusting the order, moving short but linguistically challenging constructs to higher difficulty levels. We harvested these templates semi-automatically from a corpus of real user utterances, obtained from pre-

vious data collection efforts [10]. The templates can also be manually generated if such a corpus is not available. Since not all real user utterances produce good templates, we filter them using the following constraints:

1. The sentence can be fully parsed by the English grammar
2. The Chinese translation automatically generated by the system can be fully parsed by the Chinese grammar
3. The meanings of the sentence pair, encoded as [key: value] pairs, are equivalent.

These constraints aim to ensure that the "correct answer" provided by the system can indeed be accepted by the system. Since the generated Chinese sentences are used in training the recognizer's language model, an utterance is likely to be correctly recognized and accepted whenever the user repeats a provided translation. We verified through our user study that this precaution led to improved usability of the system.

2.3. Game Level Tracking

A user-specific "start index," which controls the difficulty level of the translation task, is retained and updated from session to session. This can be viewed as a simple skill meter model [18], reflecting the student's progress in accomplishing the game task. The index defines a sliding window in the template inventory, which specifies the range of templates from which to randomly generate the task list for a game session. At the end of each session, the system decides on a new start index based on the user's performance in the session, which is represented by how many turns the user takes on average to translate each sentence (asking for help counts as a single turn). In the ideal case, a user can achieve a score of 1.0, i.e, succeeding on the first attempt for each sentence. A heuristic formula is used to update the start index as follows:

$$new\_index = old\_index + (3.0 - score) \times step\_size \quad (1)$$

In our implementation, the $step\_size$ is equivalent to the size of the sliding window, which is set to be 5% of the total number of templates. Hence, the user will stay at the same index if he/she took on average 3.0 turns per stimulus, while a "perfect" performance will advance the user by 10% of the full template space. The index decreases if the user takes more than 3 turns for each sentence. Whenever the change in the start index advances over a 10% interval, the user is notified of a "level" change (often advancing to higher levels). Many participants in our user study liked this feature and commented that it made the game fun. This observation seems to support the recent emphasis on open learner modeling (OLM): exposing the learner model to the learners could promote an individual's reflection on their evolving knowledge and on the learning process [19,20].

When a user's index remains unchanged, the task list for the next session is still very different from the previous one. This is because the "window" of templates is much larger than the number of sentences per game session, and the templates contain substitutable variables such as cities, dates and times. As a result, the game sessions are not too repetitive, even for slow-advancing users.

3. User Study

We conducted a user study involving 12 people who played the game over the Web and filled out a survey afterwards. Figure 4 displays the survey questions. The participants

Tell us about yourself:

1. Have you lived in a Chinese-speaking environment?
2. Is Chinese spoken at home? If so, which dialect?
3. How many years have you studied Mandarin formally? Informally?
4. How do you describe your Chinese proficiency?
5. Have you used other computer programs for learning Mandarin?

How do you like the translation game?

6. Was the game too easy or too difficult? In what ways?
7. Was it fun to play? Why or why not?
8. Did this game help you improve your Mandarin? If so, in what ways?
9. Would you play this game again? Why or why not?
10. Would you play a similar game which covers a different topic? If so, what topics would you like?
11. Would you recommend the system to a friend studying Mandarin? Why or why not?

How can we do better?

12. Would you prefer a more basic game which mainly focuses on vocabulary?
13. Would you prefer a more interactive game, in which the computer is a conversational partner?
14. Please tell us any suggestions which would make the system more enjoyable or effective for you.

Figure 4. Survey questions.

Subject ID	Num Utts	Max Level	WER (%)	Q6	Q7	Q8	Q9	Q10	Q11
A	859	10	9.5	E	+	+	+	+	+
B*	767	10	7.2	R	+	+	+	+	+
C*	228	10	9.8	R	+	+	+	+	+
D	174	10	12.8	E	-	+	o	+	o
E	151	4	10.0	E	+	+	+	+	+
F*	132	5	9.9	D	+	+	+	+	+
G	117	3	15.7	R	-	+	+	+	+
H	109	5	8.1	R	+	+	o	+	+
I	106	1	38.5	D	+	+	+	+	+
J*	73	7	10.3	E	+	+	+	+	+
K	71	3	17.2	D	-	+	-	-	o
L	47	1	15.1	D	o	+	+	+	+

Table 1. Statistics and survey results by user. The participants are ordered by the number of utterances they spoke with the system. The asterisk in the subject ID marks heritage speakers. See text for details.

have varied Chinese background, as indicated by their answers to survey questions 1-5. Four were considered as "heritage" speakers for whom Chinese (including dialects such as Cantonese and Shanghainese) was spoken at home. These speakers are marked by an asterisk in Table 1. Some non-heritage speakers had several years of formal and informal exposure in the past, including living in a Chinese-speaking environment for a month to a year. However, a few stated that they have not been able to keep up with the language in recent years. The most beginner participant (Subject I in Table 1) had just six months of informal exposure. About one third of the participants have used other CALL tools for Chinese, including a flash-card type of software, trial version of Rosetta Stone, and a couple of on-line Chinese dictionary Web sites.

Table 1 summarizes, for each participant, the number of Chinese utterances spoken in total, the maximum difficulty level reached, the speech recognition word error rate (WER), and the answers to survey questions 6 through 11. The WER is the number of errors made by the speech recognizer as a percentage of the number of actual words spo-

ken by a user. The participants are ordered by the number of utterances they spoke with the system. Subjects A and B are students in our group, who tested different versions of the system extensively. The heritage speakers generally have lower WERs compared to other speakers. Subject I has the highest WER, almost four times the average, which correlates with his weak prior exposure. An examination of the system log and the recorded waveforms suggests that this subject mostly imitated the system's prompted answers.

Many users provided detailed comments to the survey questions in addition to yes-no answers. We use "+" for affirmative answers and "-" for negative answers. Some users also expressed mixed opinions (e.g., "yes, if the system improves"), which were mapped to "o" in the Table. The answers to question 6 (whether the game was too easy or difficult) were represented by "D" (too difficult), "E" (too easy), and "R" (about right).

Most participants gave very positive feedback regarding the system. In particular, all users said the system was helpful (Q8 in survey). Among answers to how the system helped them improve their Chinese, the participants listed "refreshing memory on some grammar details," "learning new words and phrases," "expanding alternate phrasing," "forcing me to correct some syllable pronunciations," "learning tones," etc. Eleven out of 12 participants would play the game again, especially if the topics are expanded (Q9 and Q10). Ten of the participants would recommend the system to friends learning Chinese (Q11), the other two would after the system performance is improved.

The participants were divided on whether the game was too easy or too difficult. Four of them thought the game was too easy because "the system advances the user to higher levels too quickly," "the domain is very restricted," and "it is too easy to get help." The game was judged too difficult by Subjects I, K, and L, who also have higher WERs compared to others. A heritage speaker (F*) considered the game to be moderately challenging because of specialized vocabularies such as airline and city names. The rest of our subjects consider the difficulty level just right.

Two thirds of the users thought the game was fun to play (Q7). The other four who did not answer "yes" mentioned that persistent recognition errors were frustrating at times. One also felt that the domain was too restricted. We noted that the word error rates (an indicator of the system's performance for the user) for those four users were among the highest (surpassed only by Subject I, who has very limited Chinese exposure). This seems to suggest that the system performance is a critical factor for user enjoyment.

4. Conclusions and Future Work

In this paper we have described a Web-based voice-interactive CALL system intended to help a native speaker of English learn Mandarin through an intuitive and appealing translation game. Our user studies encourage us to believe that this can be an effective strategy for practicing a language.

In future work, we plan to investigate more sophisticated user models which can capture much richer information about the user's performance. Such information can include the particular vocabulary items or sentence constructs that the student has difficulty with (reflected by user asking for help or by increased number of unsuccessful attempts). The task list can reflect such findings to increase the student's practice in problematic areas in future game interactions.

We also plan to expand the translation game to other domains. Our subjects suggested many interesting topics, such as food, transportation, shopping, family, classroom,

etc. We envision a future version where a suite of topics are offered as options. We also plan to enhance the system's capabilities, including adding Chinese characters in the display, offering a preparation page where a learner can study new vocabularies, and adding a review stage to provide feedback on pronunciation and tones.

While our game is based on a translation task, our goal is to provide an engaging environment for the students to practice speaking the new language. We will look into other more natural ways to illicit learner's speech in the target language, for example, by using pictures. We would also like to assess whether the translation game helps students prepare for interactions with a dialogue system in the new language for booking flights.

References

[1] M. Celce-Murcia and J. Goodwin, "Teaching pronunciation," in *Teaching English as a second language*, M. Celce-Murcia, Ed. Boston: Heinle & Heinle, 1991.

[2] M. Eskenazi, "Using a computer in foreign language pronunciation training: What advantages?," *CALICO Journal*, 16(3), pp. 447–469, 1999.

[3] D. Ehsani and E. Knodt, "Speech technology in computer-aided language learnings: Strengths and limitations of a new call paradigm," *Language Learning & Technology*, 2(1), pp. 54–73, 1998.

[4] M. Eskenazi, "Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype," *Language Learning & Technology*, 2(2), pp. 62–76, 1999.

[5] S. M. Witt, *Use of Speech Recognition in Computer-assisted Language Learning*, Ph.D. thesis, Department of Engineering, University of Cambridge, Cambridge, UK, 1999.

[6] D. Hardison, "Generalization of computer-assisted prosody training: quantitative and qualitative findings," *Language Learning & Technology*, 8(1), pp. 34–52, 2004.

[7] S. Seneff, C. Wang, and J. Zhang, "Spoken conversational interaction for language learning," in *Proc. INSTIL/CALL*, 2004.

[8] W. L. Johnson, S. Marsella, and H. Vihjalmsson, "The DARWARS tactical language training system," in *Proc. Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*, 2004.

[9] A. Gruenstein, S. Seneff, and C. Wang, "Scalable and portable web-based multimodal dialogue interaction with geographical databases," in *Proc. of InterSpeech*, Pittsburgh, PA, 2006.

[10] S. Seneff and J. Polifroni, "Dialogue management in the MERCURY flight reservation system," in *Proc. ANLP-NAACL, Satellite Workshop*, Seattle, WA, 2000.

[11] C. Wang and S. Seneff, "Automatic assessment of student translations for foreign language tutoring," in *Proc. NAACL-HLT*, Rochester, NY, 2007.

[12] C. Wang and S. Seneff, "High-quality speech translation in the flight domain," in *Proc. of InterSpeech*, Pittsburgh, PA, 2006.

[13] S. Seneff, C. Wang, and J. Lee, "Combining linguistic and statistical methods for bi-directional English Chinese translation in the flight domain," in *Proc. of AMTA*, Boston, MA, 2006.

[14] J. Glass, "A probabilistic framework for segment-based speech recognition," *Computer Speech and Language*, pp. 137–152, 2003.

[15] C. Wang, D. S. Cyphers, X. Mou, J. Polifroni, S. Seneff, J. Yi, and V. Zue, "MUXING: A telephone-access Mandarin conversational system," in *Proc. ICSLP*, Beijing, China, pp. 715–718, 2000.

[16] H. C. Wang, F. Seide, C. Y. Tseng, and L. S. Lee, "MAT2000 – Design, collection, and validation on a Mandarin 2000-speaker telephone speech database," in *Proc. ICSLP*, Beijing, China, 2000.

[17] S. Seneff, "TINA: A natural language system for spoken language applications," *Computational Linguistics*, 18(1), pp. 711–714, 1992.

[18] A. T. Corbett and J. Anderson, "Knowledge tracing: Modelling the acquisition of procedural knowledge," User Modelling and User-Adapted Interaction, 4, pp. 253-278, 1995.

[19] J. Kay, "Learner know thyself: student models to give learner control and responsibility," in *Proc. of International Conference on Computers in Education*, Kuching, Malaysia, 1997.

[20] S. Bull, M. Mangat, A. Mabbott, A. S. Abu Issa and J. Marsh, "Reactions to inspectable learner models: seven year olds to university students," in *Proc. AIED Workshop on Learner Modelling for Reflection, to Support Learner Control, Metacognition and Improved Communication between Teachers and Learners*, Amsterdam, The Netherlands, 2005.