Automatic Assessment of Student Translations for Foreign Language Tutoring

Chao Wang and Stephanie Seneff
Spoken Language Systems Group
MIT Computer Science and Artificial Intelligence Laboratory
The Stata Center, 32 Vassar Street, Cambridge, MA 02139
{wangc,seneff}@csail.mit.edu

Abstract

This paper introduces the use of speech translation technology for a new type of voice-interactive Computer Aided Language Learning (CALL) application. We describe a computer game we have developed, in which the system presents sentences in a student's native language to elicit spoken translations in the target new language. A critical technology is an algorithm to automatically verify the appropriateness of the student's translation using linguistic analysis. Evaluation results are presented on the system's ability to match human judgment of the correctness of a student's translation, for a set of 1115 utterances collected from 9 learners of Mandarin Chinese translating flight domain sentences. We also demonstrate the effective use of context information to improve both recognition performance on non-native speech as well as the system's accuracy in judging the translation quality.

1 Introduction

It is widely recognized that one of the best ways to learn a foreign language is through spoken dialogue with native speakers (Ehsani and Knodt, 1998). However, this is not a practical method in the classroom setting. A potential solution to this problem is to rely on computer spoken dialogue systems to role play a tutor and/or a conversational partner. Ideally, a voice-interactive system can provide the learner with endless opportunities for practice and feedback. However, while a number of dialogue systems have been developed (or adapted) for language learning purposes (Seneff et al., 2004; Johnson et al., 2004), the issues of speech understanding of the accented and disfluent utterances of a foreign language student typically lead to unacceptable performance (Eskenazi, 1999).

A relatively successful application of speech processing technology is in the area of pronunciation training (Eskenazi, 1999; Witt, 1999; Hardison, 2004). In this case, a learner repeats words or sentences prompted by the computer, and receives feedback on the segmental and suprasegmental quality of their speech. While such systems can be useful, they do not help the student learn how to formulate sentences within the language on their own.

In this paper, we introduce a novel speech translation game which aims to provide a fun and reliable environment for language learners to gain proficiency in communication. Our application is complementary to pronunciation training in that we provide the opportunity for the students to practice expressing themselves in the new language. While we don't explicitly evaluate their pronunciations, we provide implicit feedback in that they must speak well enough for the speech recognizer to successfully process their utterances. The translation exercise can also serve as a preparation stage in which students can practice relevant vocabulary and sentence constructs, in order to prepare themselves for the more challenging task of interacting with a dialogue system in the same domain.

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 1: Example interaction between a user and the system.

Our prototype centers on the task of translating phrases and sentences from English into Chinese, in the flight reservation domain. As illustrated by the example dialogue in Figure 1, the system role plays a language tutor interacting with a Mandarin learner. The system prompts the student with randomly generated English sentences to elicit spoken Chinese translations from the learner. 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. The system keeps track of how many turns a user takes to complete all the sentences in a game session, and rewards good performance by advancing the student towards higher difficulty levels. A convenient "help" button allows the student to request a translation of the current game sentence, to help them overcome gaps in their knowledge of the linguistic constructs or the vocabulary. The student can also type any English sentences within the domain to obtain a reference translation. The system utilizes an interlingua-based bidirectional translation capability, described in detail in (Wang and Seneff, 2006; Seneff et al., 2006). Both Chinese and English sentences are parsed into a common meaning representation, which we loosely refer to as an "interlingua," from which paraphrases in both languages can be automatically generated using formal generation rules.

The key to a successful tutoring system lies in its ability to provide immediate and pertinent feedback on the student's performance, similar to a human tutor. A central focus of this paper is to address the challenging problem of automatically assessing the appropriateness of a student's translation. At first glance, our task appears to share much in common with machine translation (MT) evaluation (Hovy et al., 2002). Indeed, both are trying to assess the quality of the translation output, whether it is produced by a computer or by a foreign language student. Nevertheless, there also exist several fundamental distinctions. Automatic MT evaluation methods, as represented by the well-known Bleu metric (Papineni et al., 2001), assume the availability of human reference translations. The algorithms typically compare MT outputs with reference translations with the goal of producing a quality indicator (on a numeric scale) that correlates with human judgement. In contrast, our algorithm operates in the absence of human generated reference translations[1]. Furthermore, our application requires the evaluation algorithm to make accept/reject decisions on each *individual* translation, in the same way as a language tutor determines whether a translation is acceptable or not. While our task is more demanding, it is made possible by operating in restricted domains.

The remainder of the paper is organized as follows. In Section 2, we present an interlingua-based approach for verifying the correctness of the student's spoken translation. Section 3 describes the

[1]We employ a grammar of recursive rewrite rules to generate a very large number of English prompt sentences. It would be too costly and time-consuming to generate human translations to cover this space.

evaluation framework, followed by results and discussions in Section 4. Finally, we discuss future plans for extending our work.

2 Methodology

The two most important aspects in the human evaluation of translation quality are *fluency* and *fidelity* (Hovy et al., 2002). In our case, we consider a student's translation to be acceptable if it is well-formed (high fluency) and conveys the same meaning as the input sentence (high fidelity). We designed our interlingua-based evaluation algorithm following these two principles. The algorithm uses parsability to verify fluency. Fidelity is examined by extracting and comparing semantic information from the translation pairs. In the following, we begin by describing the basic steps involved in our translation verification algorithm. We then discuss different strategies for integrating with the speech recognition system.

2.1 Parsing

Our framework depends strongly on an ability to parse both the English and Chinese sentences into a common interlingual meaning representation. Parsing is critical both for producing the two paraphrases of the student's utterance and for judging the quality of their provided translation. Both English and Chinese grammars are needed to analyze the source and target sides of each translation pair. The grammars have been carefully constructed so that meaning representations derived from both languages are as similar as feasible.

We utilized a parser (Seneff, 1992) that is based on an enhanced probabilistic context-free grammar (PCFG), which captures dependencies beyond context-free rules by conditioning on the external left-context parse categories when predicting the first child of each parent node. While we use a specific grammar for analyzing flight domain sentences, we emphasize domain portability of the grammar by using mainly syntactic information in the majority of the parse tree rules. Semantics are introduced near the terminals, mostly involving adjectives, verbs, nouns and proper noun classes. Rules for general semantic concepts such as dates and times are organized into sub-grammars that are easily embedded into any domain. We have successfully applied the same strategy in developing both the Chinese and English grammars. Once a parse tree is obtained, selected parse categories are extracted to form a hierarchical meaning representation encoding both syntactic and semantic information.

2.2 Semantic Information Comparison

In principle, we can directly compare the meaning representations derived from the source and target sides of the translation pair to determine their equivalence. In practice, the meaning representation still captures too much language-specific detail, which makes the comparison prone to failure. Even the pair of English utterances, "How much is the second flight?" and "What is the price of the second flight?" have essentially the same meaning, but would not produce identical meaning representations. Across languages, this situation becomes much worse.

We adopted two complementary strategies to increase the chance of a match between the English prompt and the student translation. First, the English prompt is translated into a reference Chinese translation using the existing interlingua translation capability. This extra step aims at reducing discrepancies caused by syntactic structure differences between the two languages. Secondly, we abstract from the original meaning representation into a simple encoding of key-value (KV) pairs. This is accomplished using a language generation system (Baptist and Seneff, 2000), with generation rules determining what information to extract from the original hierarchical meaning representation. Figure 2 shows a couple of examples of the KV representation that we used for scoring.

Another important role of the KV generation step is to bring in a flexible mechanism for defining equivalence, which is a tricky task even for human evaluators. For example, while it is somewhat obvious that "(1) Give me flights leaving around nine p m" is equivalent to "(2) Give me flights departing around nine p m," it is unclear whether these two sentences are equivalent to "(3) Give me flights around nine p m" or even "(4) I would like to leave around nine p m." From a pragmatic point of view, the same speaker intention can be inferred from the four sentences. On the other hand, it can be argued that (1) and (2) are completely interchangeable

```
{c eform
   :topic  "fare"
   :airline_name "united"
   :trace "how much" }
```

lian2 he2 hang2 kong1 de5 piao4 jia4 shi4 duo1 shao3?
(*What is the fare on United airlines?*)

```
{c eform
   :topic "flight"
   :source "paris"
   :departure_time {c eform
                      :clock_hour 9
                      :xm "a m" }}
```

shang4 wu3 jiu3 dian3 cong2 ba1 li2 chu1 fa1 de5 ban1 ji1.
(*Flights from Paris leaving at nine o'clock in the morning.*)

Figure 2: Frame representation of the key-value information for two example Chinese sentences.

while (3) and (4) could not substitute for (1) or (2) in some circumstances. Criteria for equivalence can be controlled by what is extracted from the meaning representation. If only a `departure_time` key is generated for the sentences, then all four sentences will be equivalent. However, if more information is preserved in the KV pairs, for example, a `topic` key with value `flight`, then sentence (4) will not be considered as equivalent to sentences (1)-(3). Considering that our intended application is language tutoring, we lean towards a stricter criterion for defining equivalence. The KV generation rules are developed manually, guided by human-rated development data. The KV inventory includes over 80 unique keys.

Once the KV pairs are obtained from the prompt (reference) and the student translation (hypothesis), a recursive procedure is applied to compare all the keys in the reference and hypothesis KV frames. Mismatches are tabulated into substitutions (different values for the same key), deletions (extra keys in the reference), and insertions (extra keys in the hypothesis). A perfect match is achieved if there are no mismatch errors. Figure 3 summarizes the procedure to evaluate students' spoken translations.

Partial match for a good student translation is a common problem caused by speech recognition errors, particularly on dates and times. It is natural for the student to just repeat the "incorrect" piece after noticing the error in the system's paraphrases. Hence, in the tutoring application, we added a submatch mode to the comparison algorithm, which works in a divide-and-conquer manner. All matching KV pairs in each turn are checked off from the reference, and a subsequent submatch succeeds once there are no remaining KV pairs unaccounted for. One limitation of the incremental comparison algorithm is that it ignores insertion errors. The tutoring system provides a special reply message when a sentence is translated via partial matches accomplished over a series of utterances, to distinguish from the case of a perfect match in a single turn, as illustrated in the example dialogue.

2.3 Integration with Speech Recognition

A user's utterance is first processed by the speech recognizer to produce word hypotheses. The recognizer is configured from a segment-based speech recognition system (Glass, 2003), using Chinese acoustic models trained on native speakers' data (Wang et al., 2000a; Wang et al., 2000b). Tone features are ignored in the acoustic models; however, the language model implicitly captures some tone constraints. This is preferred over modeling tone explicitly, considering that non-native speakers typically make many tone errors. The language model was initially trained on Chinese translations of English sentences generated from the templates used in the game, and later augmented with additional data collected from users. The recognizer can output multiple hypotheses in the form of an *N*-best list. The parser is able to convert the *N*-best list into a lattice, and re-select a best hypothesis based on a combination of recognition and parsing scores.

Poor recognition on non-native speech is a major performance issue for CALL application. In our domain, dates, times, and flight numbers are particularly challenging entities for the recognizer. Recognition error typically results in false rejection, causing frustration to the user. Since the system has explicit knowledge of the sentence the student is trying to produce, it should be feasible to exploit this knowledge to improve speech understanding. A plausible strategy is to dynamically adjust the recognizer's language model in anticipation of what the user is likely to say, as exemplified by dialogue context dependent language models (Solsona et al., 2002).

In theory, we could use the automatically gener-

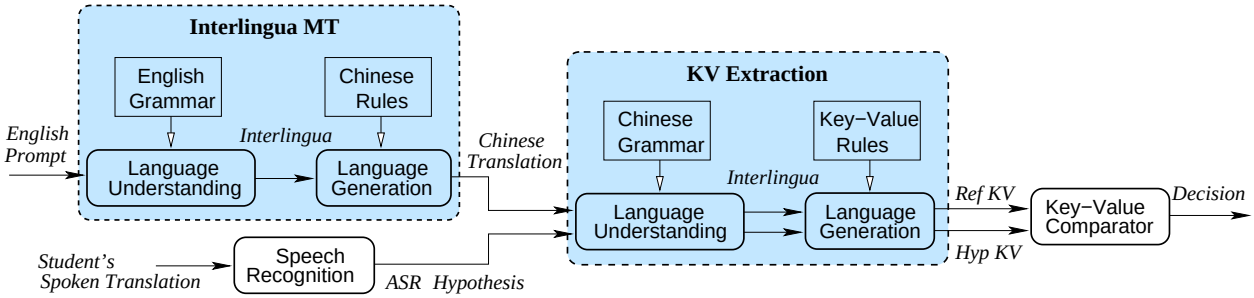

Figure 3: Schematic of procedure to evaluate students' spoken translations.

ated reference translation to explicitly bias the language model. However, one has to take care not to bias towards the correct response so strongly that the student is allowed to make mistakes with impunity. Furthermore, this strategy would not generalize to cover all the possible legitimate translations a student might produce for that prompt. Instead, we devised a simple strategy that overcomes these issues. We select a preferred hypothesis from the *N*-best list if its KV representation matches the reference. Thus the student has to speak well enough for a correct answer to appear somewhere in the N-best list, without any manipulations of the recognizer's language model. If the parser fails to find a perfect match in the *N*-best list, it will choose the hypothesis with the best score, or fall back to the recognizer's top hypothesis if no parse theory could succeed.

3 Evaluation Framework

Given a translation pair, the goal of our algorithm is to make the same accept/reject decision as a human evaluator. Hence, we can evaluate our algorithm in a classification framework. In this section, we first present the data collection and labeling effort. We then describe a baseline system based on a variant of the Bleu metric. Finally we briefly describe the metrics we used to evaluate our algorithms.

3.1 Data Collection and Labeling

During the course of developing a prototype game system, two developers and two student testers interacted extensively with the system. A total of 2527 Chinese waveforms, recorded during this process, became development data for finding gaps in the interlingua-based matching method and for tuning parameters for the baseline method.

For evaluation, we use 1115 utterances collected from 9 users with varying degrees of Chinese exposure. These subjects were asked to play the translation game over the Web and fill out a survey afterwards. They came from a rich background of Chinese exposure, include advanced "heritage" speakers of Chinese (including dialects such as Cantonese and Shanghainese), as well as novices who just completed two semesters of a college-level Chinese class.

The speech waveforms recorded from the interactions were manually transcribed with orthography, gender, and speaker information. The transcriber was instructed to transcribe spontaneous speech effects, such as false starts and filled pauses. However, tonal mispronunciations are completely ignored, and segmental errors are largely ignored to the extent that they do not result in a different syllable.

The translation pairs (the English prompt and the orthographic transcription of the student translation) were rated independently by two bilingual speakers to provide reference labels for evaluating the verification algorithm. The two raters, both native in Chinese and fluent in English, labelled each translation with either an "accept" or a "reject" label. Translations can be rejected because of bad language usage (including false starts) or because of mismatches in meaning. One labeller rated both development and test data, while the second labeller only rated the test data. The interlabeller agreement on the test data has a kappa score (Uebersax, 1998) of 0.85. The subset of data for which there was disagreement were relabelled by the two raters jointly to reach a consensus.

3.2 Baseline

The Bleu metric has been widely accepted as an effective means to automatically evaluate the quality of machine translation outputs (Papineni et al.,

2001). An interesting question is whether it would be useful for the purpose of assessing the appropriateness of translations produced by *non-native speakers* at a sentence by sentence granularity level. We developed a simple baseline algorithm using the NIST score, which is a slight variation of Bleu[2]. Given an English prompt, the interlingua-based machine translation system first produces a reference translation. The student's translation is then compared against the machine output to obtain a NIST score. The translation is accepted if the score exceeds a certain threshold optimized on the development data.

Figure 4 plots the Receiver Operating Characteristics (ROC) curve of the baseline algorithm, obtained by varying the NIST score acceptance threshold. Each point on the curve represents a tradeoff between accepting an erroneous translation (False Accept) and rejecting a good one (False Reject). As shown in the plot, the NIST score based ROC curve is far from reaching the ideal top-left corner. For language tutoring purposes, it is desirable to operate in the low false acceptance region. However, a 20% false acceptance rate will result in the system rejecting over 35% of correct student translations. The operating point that minimizes overall classification error turns out to be biased towards leniency, falsely accepting over 60% of translations that are rejected by human raters. The resulting minimum error rate on development data transcripts is 23.0%, with a NIST score threshold of 3.16. The threshold for automatic speech recognition (ASR) outputs was optimized separately using the 1-best hypotheses of utterances in the development data. The optimal threshold on ASR outputs is 1.60, resulting in a classification error rate of 24.1%. The majority classifier, corresponding to the $(1, 1)$ point on the curve, translates into a 31.6% error rate on the development data.

3.3 Evaluation Metrics

We evaluated the overall system performance on test data using human decisions as ground truth. Although we can not generate an ROC curve for our proposed algorithm (because it is a non-parametric method), we plot its performance along with the ROC curve of the baseline system for a more thorough comparison.

[2]We determined empirically that the NIST score works slightly better than the Bleu score in our application. The scores are computed using the NIST MT scoring tool from: `ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v11b.pl`

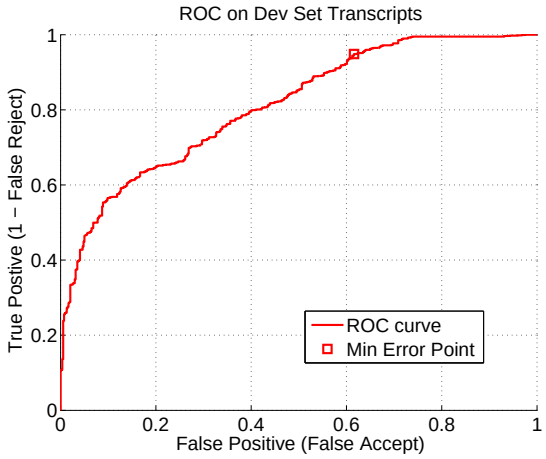

Figure 4: ROC curve by changing acceptance threshold on the NIST score on transcriptions of development data.

We evaluated the different ASR integration strategies (1-best hypothesis, 10-best hypotheses, using contextual constraints from reference KV) based on sentence classification error rates as well as speech recognition performance.

4 Results and Discussions

Table 1 summarizes the false accept, false reject, and overall classification error rates on unseen test data. With manual transcripts as inputs, the baseline algorithm using the NIST score achieved a classification error rate of 19.3%, as compared with 25.0% for the trivial case of always accepting the user sentence (Majority classifier). The KV-based algorithm achieved a much better performance, with only a 7.1% classification error rate. This translates into a kappa score of 0.86, which is slightly above the level of agreement initially achieved by the two labellers. Note that the performance difference compared to the baseline system is mostly attributed to a large reduction in the "False Accept" category.

Interestingly, the NIST method degrades only slightly when it is applied to the speech recognition 1-best output rather than the transcript. However, this result is deceptive, as it is now even more bi-

Transcript	False Reject	False Accept	Classification Error
Majority	0.0%	100%	25.0%
NIST	8.0%	54.5%	19.6%
KV	7.3%	6.8%	7.2%

ASR	False Reject	False Accept	Classification Error
NIST	4.2%	77.1%	22.4%
KV 1-best	32.1%	4.3%	25.1%
KV 10-best	27.0%	7.2%	22.1%
KV Context	13.5%	14.7%	13.8%

Table 1: Classification results for various evaluation systems, on both transcripts and automatic speech recognition (ASR) outputs. Note that the "KV Context" condition favors a hypothesis that matches the prompt KV.

ased towards a "False Accept" strategy, causing over three quarters of the students' erroneous utterances to be accepted. The KV method is much more susceptible to speech recognition error because of its deep linguistic analysis. For instance, any recognition errors causing a parse failure will result in a "reject" decision, which explains the high error rate when only the 1-best hypothesis is used. However, the KV algorithm can improve substantially by searching the full N-best list (N = 10) for a plausible analysis. When contextual information (KV Context) is used, our simple strategy of favoring the hypothesis matching the reference KV reduces the classification error rate dramatically.

A plot of the receiver operating characteristics of these methods in Figure 5 reveals a clear picture of the performance differences. All of the KV points are clustered in the upper left corner of the plot, above the ROC curve of the NIST-based method. The NIST-score based classifier (represented by the square marker on the ROC curve) is heavily biased towards making the acceptance decision (the majority class). In contrast, the KV method operates in the low "False Accept" area. It achieves a much lower false rejection rate when compared with the NIST method operating at an equivalent false acceptance point.

Although the classification error rate clearly im-

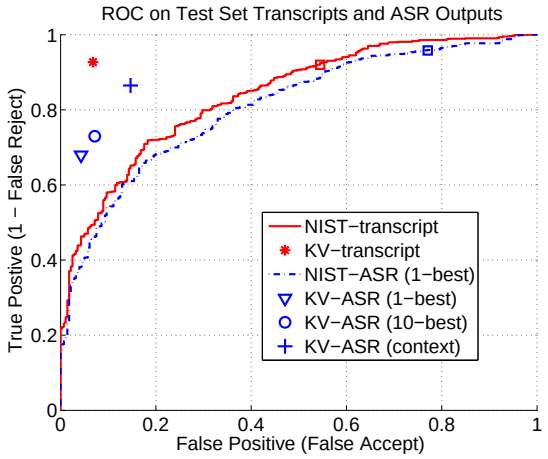

Figure 5: Comparison of ROC of different methods.

	Syllable		Sentence	
	ER(%)	RR(%)	ER(%)	RR(%)
1-best	11.6	-	40.4	-
10-best	10.7	7.8	38.7	4.2
Context	8.7	25.0	30.0	25.7

Table 2: Comparison of speech recognition performance in syllable error rates and sentence error rates, for three different strategies of utterance selection from an N-best list. (ER stands for error rate, RR stands for relative reduction.)

proves when the KV method makes use of the N-best list and incorporates contextual constraints, the ROC plot seems to suggest that the error reduction might simply be attributed to a shift in the operating point: the improvements are caused by a bias towards making the majority class decision. We use improvements in speech recognition to demonstrate that this is not the case (at least not entirely). Table 2 summarizes the syllable and sentence error rates on the test data, for the three configurations discussed previously (1-best, 10-best, and Context). By using a tighter integration with the parser with contextual constraints, we greatly improved speech recognition performance, marked by reductions of syllable and sentence error rates by 25% and 25.7% respectively.

5 Conclusions and Future Work

In this paper, we have presented an algorithm for automatically assessing spoken translations produced by language learners. The evaluation results demon-

strated that our method involving deep linguistic analysis of the translation pair can achieve high consistency with human decisions, and our strategy of incorporating contextual constraints is effective at improving speech recognition on non-native speech. While our solution is domain specific, we emphasize domain portability in the linguistic analysis modules, so that similar capabilities in other domains can be quickly developed even in the absence of training data. Our interlingua framework also makes the methodology agnostic to the direction of source and target languages. A similar application for native Mandarin speakers learning English could be instantiated by using the same components for linguistic analysis.

A major challenge in our problem is in determining equivalence between the meanings of a translation pair. While our approach of using a rule-based generation system gives the developer great flexibility in deriving an appropriate KV representation, the comparison algorithm is somewhat primitive: it relies entirely on the generation rules to produce the right KV representation. In future work, we plan to apply machine learning techniques to this problem. With the data we have collected and labelled (and the effort is ongoing), it becomes feasible to examine the use of data-driven methods. As alluded to in our evaluation methodology, we can cast the problem into a classification framework. Lexical, *n*-gram, and alignment based features can be extracted from the translation pairs, which can be further enhanced by features obtained from deep linguistic analysis. This will relieve the burden on the semantic analysis component, and improve the overall portability of our approach.

We also plan to expand our application to many other domains appropriate for language learning, and test the effectiveness of the translation game as a means for language learning.

6 Acknowledgements

This research is supported in part by ITRI and the Cambridge MIT Initiative. The authors would like to acknowledge Yushi Xu for annotating the data. We are also grateful to Michael Collins and the anonymous reviews for their helpful comments and suggestions.

References

L. Baptist, S. Seneff. 2000. Genesis-II: A versatile system for language generation in conversational system applications. In *Proc. ICSLP*, Beijing, China.

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

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

J. Glass. 2003. A probabilistic framework for segment-based speech recognition. *Computer Speech and Language*, 17:137–152.

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

E. Hovy, M. King, A. Popescu-Belis. 2002. Principles of context-based machine translation evaluation. *Machine Translation*, 7(1):43–75.

W. L. Johnson, S. Marsella, H. Vihjalmsson. 2004. The DARWARS tactical language training system. In *Proc. I/ITSEC*.

K. Papineni, S. Roukos, T. Ward, W.-J. Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*.

S. Seneff, C. Wang, J. Zhang. 2004. Spoken conversational interaction for language learning. In *Proc. INSTIL/CALL*.

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

S. Seneff. 1992. TINA: A natural language system for spoken language applications. *Computational Linguistics*, 18(1).

R. A. Solsona, E. Fosler-Lussier, H.-K. J. Kuo, A. Potamianos, I. Zitouni. 2002. Adaptive language models for spoken dialogue systems. In *ICASSP*.

J. S. Uebersax. 1998. Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin*, 101:140–146.

C. Wang, S. Seneff. 2006. High-quality speech translation in the flight domain. In *Proc. of InterSpeech*.

C. Wang, D. S. Cyphers, X. Mou, J. Polifroni, S. Seneff, J. Yi, V. Zue. 2000a. MUXING: A telephone-access Mandarin conversational system. In *Proc. ICSLP*, 715–718, Beijing, China.

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

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