INTENTION-BASED CORRECTIVE FEEDBACK GENERATION USING CONTEXT-AWARE MODEL

Sungjin Lee, Cheongjae Lee, Jonghoon Lee, Hyungjong Noh, and Gary Geunbae Lee
Pohang University of Science and Technology (POSTECH), Korea
{junion, lcj80, jh21983, nohhj, gblee} @postech.ac.kr

Keywords: Dialog-based Computer Assisted Language Learning, Dialog System, Conversational Tutoring

Abstract: In order to facilitate language acquisition, when language learners speak incomprehensible utterances, a Dialog-based Computer Assisted Language Learning (DB-CALL) system should provide matching fluent utterances by inferring the actual learner's intention both from the utterance itself and from the dialog context as human tutors do. We propose a hybrid inference model that allows a practical and principled way of separating the utterance model and the dialog context model so that only the utterance model needs to be adjusted for each fluency level. Also, we propose a feedback generation method that provides native-like utterances by searching Example Expression Database using the inferred intention. In experiments, our hybrid model outperformed the utterance only model. Also, from the increased dialog completion rate, we can conclude that our method is suitable to produce appropriate feedback even when the learner's utterances are highly incomprehensible. This is because the dialog context model effectively confines candidate intentions within the given context.

1 INTRODUCTION

Second language acquisition (SLA) researchers have claimed that feedback provided during conversational interaction facilitates the acquisition process (Long, 2005; Swain, 1996). Helpful interactional processes include the *negotiation of meaning* and provision of *recasts*, both of which can supply corrective feedback to let learners know that their utterances were problematic. A further interactional process that can result from feedback is known as *modified output*. For example, consider the interactional processes, in which the system negotiates to determine the meaning using a clarification request in response to the learner's unnatural expression (Table 1). The language learner modified the original utterance to convey the intended meaning by referring to the recast provided by the system.

Unfortunately, conversational interaction is one of the most expensive ways to teach a language. Thus, interest in developing Dialog-based Computer Assisted Language Learning (DB-CALL) systems is rapidly increasing. However, just using conventional dialog systems in a foreign language would not be beneficial because language learners commit numerous and diverse errors. A DB-CALL system should be able to understand language learners' utterances in spite of these obstacles. Also, it is desirable to offer appropriate feedback.

To achieve this goal, rule-based systems usually anticipate error types and hand-craft a large number of error rules but this approach makes these methods sensitive to unexpected errors and diverse error combinations (Schneider and McCoy, 1998; Morton and Jack, 2005; Raux and Eskenazi, 2004). A more serious problem is that just correcting grammatical errors cannot guarantee that the utterance is fluent and meaningful. Therefore, we argue that the proper

Table 1: An example dialog in which the DB-CALL system returns a feedback recommending use of a native-like utterance

Speaker	Intention	Dialog
System:	*wh-question (trip-purpose)*	What is the purpose of your trip?
User:	*inform (trip-purpose)*	My purpose business
System:	*clarify (under-standing)*	Sorry, I don't understand. What did you say? ← **Clarification request** ***On screen****: try this expression "I am here on business"* ← **Recast**
User:	*inform (trip-purpose)*	I am here on business ← **Modified output**

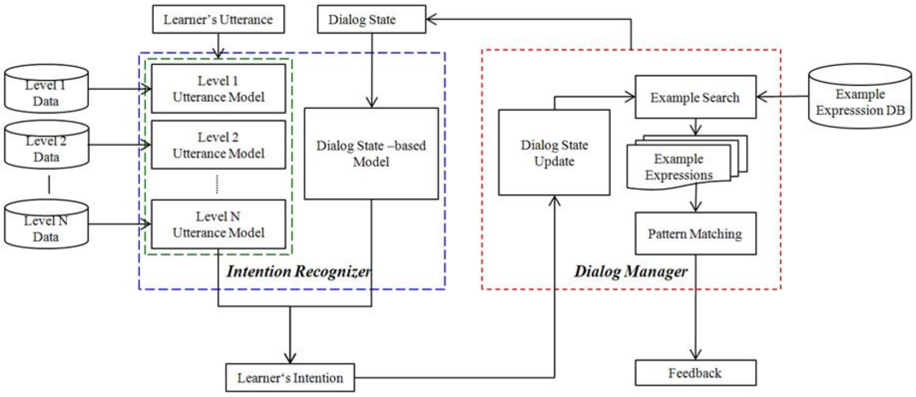

Figure 1: System architecture

language tutoring methodology is not to correct specific errors but to provide native-like utterance examples which realize the user's intention.

To accomplish this purpose, as human tutors do, we first infer the actual learners' intention from the erroneous utterances by taking not only the utterance itself but also the dialog context into consideration, and then generate a corrective feedback based on the inferred intention.

The remainder of this paper is structured as follows. Section 2 briefly describes related studies. Section 3 introduces the system architecture and operation. Section 4 presents the detailed description of our hybrid intention recognition model. Section 5 describes the experimental results to assess the method's potential usefulness. Finally, Section 6 gives our conclusion.

2 RELATED WORKS

There are several studies on general dialog systems which have examined incorporating the dialog context into recognizing dialog acts. Due to the difficulties of extracting and employing rich dialog context, most of them included just a few types of context such as previous dialog act (Poesio and Mikheev, 1998), or dialog state in finite-state model (Bohus and Rudnicky, 2003). Recently, Ai et. al. (2007) investigated the effect of using rich dialog context and showed promising results. The ways to incorporate the dialog context mostly involved just combining all features both from the utterance and the context into one feature set which was then used to train inference models. For DB-CALL, however, such approaches can be problematic, because distinct handling for each of the fluency level is important in a language learning setting. Given a dialog scenario, the dialog context model is relatively invariant; thus we propose a hybrid model that combines the utterance model and the dialog context model in a factored form. This approach allows us to adjust the hybrid model to a required fluency level by replacing only the utterance model.

3 SYSTEM ARCHITECUTRE AND OPERATION

The whole system consists of the intention recognizer and the dialog manager (Fig. 1). The intention recognizer is a hybrid model of the dialog state model and one of the utterance models. A specific utterance model is chosen according to a learner's proficiency level. When the learner utters, the utterance model elicits n-best hypotheses of the learner's intention, and then they are re-ranked by the results of the dialog state model. The detailed algorithm will be described at the next section.

The role of the dialog manager is to generate system responses according to the learner's intention and also generate corrective feedback if needed. Corrective feedback generation takes two steps: 1) Example Search: the dialog manager retrieves example expressions by querying Example Expression Database (EED) using the learner's intention as the search key. 2) Example Selection: The dialog manager selects the best example which maximizes the similarity to the learner's utterance based on lexico-semantic pattern matching.

If the example expression is not equal to the learner's utterance, the dialog manager shows the example as recast feedback and conduct a clarification request to induce learners to modify their utterance (Table 1). Otherwise, the dialog manager shows one of the retrieved examples as paraphrase feedback so that learners may acquire another expression with the same meaning. Sometimes, students have no idea about what to say and they cannot continue the dialog. In such a case, time out occurs and the utterance model does not

Table 2: Representation of dialog context and an example for immigration domain

Dialog Context Features	
PREV_SYS_INT	Previous system intention Ex) PREV_SYS_INT = wh-question(job)
PREV_USR_INT	Previous user intention Ex) PREV_USR_INT = inform(job)
SYS_INT	Current system intention Ex) SYS_INT = confirm(job)
INFO_EX_STAT	A list of exchanged information states which is essential to successful task completion; (c) denotes *confirmed*, (u) *unconfirmed* Ex) INFO_EX_STAT = [nationality(c), job(u)]
DB_RES_NUM	Number of database query results Ex) DB_RES_NUM = 0

generate hypotheses. Hence, the dialog system searches EED with only the result of the dialog state model and shows the retrieved expression as suggestion feedback so that students can use it to continue a conversation

4 HYBRID INTENTION RECOGNITION MODEL

Our representation of user intention consists of dialog act and type of subtask as shown in Table 1. For example, the first system utterance *"What is the purpose of your trip?"* can be abstracted by the intention *wh-question (trip-purpose).*

The hybrid model merges hypotheses from the utterance model with hypotheses from the dialog context model to find the best overall matching user intention. In the language production process, user intentions are first derived from the dialog context; subsequently the user intentions determine utterances (Carroll, 2003). By using this dependency and the chain rule, the most likely expected user intention $I(D,U)$ given the dialog context D and the utterance Ucan be stated as follows:

$$I(D,U) = \underset{I}{argmax}\, P(I|D,U) \quad (1)$$

$$= \underset{I}{argmax} \frac{P(D,I,U)}{P(D,U)} \quad (2)$$

$$= \underset{I}{argmax} \frac{P(D)P(I|D)P(U|I)}{P(D,U)} \quad (3)$$

By using Bayes' rule, Eq. (3) can be reformulated as:

$$I(D,U) = \underset{I}{argmax} \frac{P(D)P(I|D)P(I|U)}{P(D,U)P(I)} \quad (4)$$

P(D) and *P(D, U)* can be ignored, because they are constant for all *I* (Eq. 5):

$$I(D,U) = \underset{I}{argmax} \frac{P(I|D)P(I|U)}{P(I)} \quad (5)$$

In this formula, *P(I|D)* represents the dialog context model and *P(I|U)* represents the utterance model. The next two subsections discuss each sub-model in detail.

4.1 Utterance Model

To predict the user intention from the utterance itself, we use maximum entropy model (Ratnaparkhi, 1998) trained on linguistically-motivated features. This model offers a clean way to combine diverse pieces of linguistic information. We use the following linguistic features for the utterance model.

- *Lexical word features*: Lexical word features consist of lexical tri-grams using current, previous, and next lexical words. They are important features, but the lexical words appearing in training data are limited, so data sparseness problem can arise.
- *POS tag features*: POS tag features also include POS tag tri-grams matching the lexical features. POS tag features provide generalization power over the lexical features.

The objective of this modeling is to find the *I* that maximizes the conditional probability, *P(I|U)* in Eq. (5), which is estimated using Eq. (6):

$$P(I|U) = \frac{1}{Z} exp\left(\sum_{k=1}^{K} \lambda_k f_k(I,U)\right) \quad (6)$$

where K is the number of features, f_k denotes the features, λ_k the weighted parameters for features, and Z is a normalization factor to ensure $\Sigma P(I|U) = 1$. We use a limited memory version of the quasi-Newton method (L-BFGS) to optimize the objective function.

4.2 Dialog Context Model

Our representation of a dialog context consists of diverse pieces of discourse and subtask information as shown in Table 2. The task of predicting the probable user intention in a given dialog context can be viewed as searching for dialog contexts that are similar to the current one in dialog context space and then inferring to the expected user intention from the user intentions of the dialog contexts found.

Therefore, we can formulate the task as the k-nearest neighbors (KNN) problem (Dasarathy, 1990). We had a number of reasons for choosing instance-based learning methodology. First, instance-based learning provides high controllability for tuning the model incrementally during operation, which is practically very desirable property. Second, an elaborate similarity function can be applied. Many of other approaches, e.g. maximum entropy model used in the utterance model, express the similarity between states in a simplest manner through the features that the states share, losing elaborate regularities between features. For the dialog context model, we can easily predict which features become important features to measure similarity conditioning on certain values of other features using general discourse knowledge. For example, if the current system dialog act is "inform", the number of database query results becomes an important feature. If the number of results is greater than one, the most likely expected user intention would be "select". If the number of results equals one, "ack" would be the most probable intention. Otherwise, the users might want to modify their requirements. Another example, if all exchanges of information are confirmed and the current system intention is "wh-question", the current system intention itself becomes the most important feature to determine the next user intention.

However, the conventional KNN model has two drawbacks. First, it considers no longer the degree of similarity after selecting k nearest contexts, hence intentions that occur rarely cannot have a chance to be chosen regardless of how close they are to the given dialog context. The second drawback is that if dialog contexts with, say, intention A, are locally condensed rather than widely distributed, then A is specifically fitted intention to the local region of the dialog context. So the intention A should be given greater preference than other intentions. To cope with these drawbacks, we introduce a new concept, locality, and take both similarity and locality into account in estimating the probability distribution of the dialog context model (Eq. 9, 10).

The similarity function is defined as the following equation:

$$Similarity(D, D') = \sum_{k=1}^{K} \lambda_k f_k(D, D') \tag{7}$$

where K is the number of features, f_k denotes the feature functions, λ_k the weighted parameters for features. Our feature functions first include the simplest tests, whether a feature is shared or not, for each feature of a dialog context (Table 2). For composite features, individual tests are also included for each constituent to alleviate data sparseness problems. For example, we include feature functions not only for system intention but also for its constituents, system dialog act and type of subtask. In addition, we include a number of feature functions which test the elaborate rules as illustrated in the examples above. The weighted parameters are given initial values based on general discourse and task knowledge and optimized on the development data set with minimum error rate criteria.

The locality function is the ratio between the number of elements of the set $S_{I,D}$ and the number of elements of the set S_I:

$$Locality(I, D) = \frac{n(S_{I,D})}{n(S_I)}, \tag{8}$$

where $S_I = \{D' \mid \forall\, D'\, intention(D') = I\}$, and $S_{I,D} = \{D' \mid \forall\, D' \in S_D\ and\ intention(D') = I\}$, and S_D is the set of k nearest neighbors of the given dialog context D.

The score function calculates the score of the intention I based on the set of k nearest dialog contexts using both similarity and locality:

$$\begin{aligned} &Score(I, D) \\ &= Locality(I, D) \sum_{D_i \in S_{I,D}} Similarity(D_i, D) \end{aligned} \tag{9}$$

To let the dialog context model be a probability distribution, the score function is divided by the normalization factor:

$$P(I|D) = \frac{Score(I, D)}{\Sigma_{I'} Score(I', D)} \tag{10}$$

5 EXPERIMENTS

To verify the proposed model, we conducted a case study for dialogs in a system devoted to immigration into an English-speaking country. We used the example based dialog management method (Lee et. al., 2009) to build the dialog system. We used the dialog corpus in an immigration domain which consists of 192 dialogs (3,517 utterances, 18.32 utterances per dialog). The number of types of dialog act is seven and the number of types of subtask is eleven. We annotated each utterance in dialogs by two steps. In the first step, we manually annotated each utterance with the speaker's intention and component slot-values. In the second step, we automatically annotated each utterance with the discourse information (Table 2) using in-house software.

5.1 An example dialog

The following dialog gives a flavor of the system in real action. A brief commentary is given at key points in the dialog and is printed in italics.

S1: Hi. This is a dialogue system for English learning. May I ask your name for authentication?

U1: My name is Hong Kil Dong

S2: You are a new user. We have four themes such as immigration, transportation, hotel, and restaurant theme. Which theme do you want to learn?

U2: I'd like the immigrant theme

S3: The immigrant theme is selected. Let's get started

S4: Hello

U3: Hi

S5: Your passport please

U4: Here this

***On screen:** [Recast] Here you are*

[Recast] is an example utterance provided by recast feedback. "Here this", the student's utterance, is understandable based on context but it is not a good expression. The system recommends a better expression so that the student becomes better in a next dialogue.

S6: Sorry, I don't understand. What did you say?

U5: Here you are

S7: What's your nationality?

U6: I am a Korean

***On screen:** [Paraphrase] I come from Korea*

When the user utterance is correct the system shows another possible expression by paraphrase feedback.

S8: What's the purpose of your visit?

U7: Business

***On screen:** [Paraphrase] I am here on business*

If a user input is simple words or a phrase, then the system shows a complete sentence.

S9: How long are you staying?

U8: I stay for three weeks

***On screen:** [Recast] I will stay for three weeks*

The user utterance is semantically understandable but it is grammatically wrong. The system shows a grammatically correct sentence on screen which is obtained by searchingEED.

S10: Sorry, I don't understand. What did you say?

U9: I will stay for three weeks

S11: Anything to declare?

U10: Pardon me?

S12: Anything to declare?

U11: Nothing

***On screen:** [Paraphrase] Just my personal things*

S13: What kind of job?

U12: [TIMEOUT]

***On screen:** [Suggestion] I am a student*

[Suggestion] is an example utterance provided by suggestion feedback. The system knows that it asked the user about a job in the previous question. And it also knows that it already got information about nationality, purpose of visit and period through the dialog history. Therefore by using such dialogue history, the system can find out a proper example from Example Expression DB.

U13: I am a computer programmer

S14: Ok, no problem

U14: Thank you. Bye

S15: The immigrant theme has been done. Your total score is B. In next time, you better use more comprehensive and complete expressions, and try to say more quickly. Would you like to continue learning another theme?

The system provides assessment feedback which includes total score and specific messages that inform the student which part they need to improve.

U15: No.

S16: Study session ends.

Figure 2: An example of real conversation in an immigration domain

5.2 Simulated language learner

In this experiment, instead of involving real language learners, we simulated them by injecting grammar errors into clear utterances generated using the user simulation method described in (Jung et. al., 2009). We did this because we want to freely adjust the fluency level of users for intensive tests in a variety of settings, and because the cost of engaging large numbers of language learners with diverse fluency levels during development is particularly high. Employing a simulated user will greatly accelerate the development cycles.

To simulate various fluency levels, we examined the Korean Student English Essay (KSEE) corpus which is a collection of essays by Korean university students. The corpus consists of 119 essays (915 sentences). We annotated the corpus with the error tags defined in (Michaud, 2002). The frequencies of error types were measured. In total, 65 error types and 2,404 instantiated errors were discovered. We classified error types into three categories:

Table 3: Three categories of error types and the top 5 error types in each category

Category	Error type with category
Substitution (47%)	Spell (71%) Plural Form (14%) Subject Verb Agreement (10%) Incorrect Preposition (3%) Incorrect Determiner (2%)
Deletion (36%)	Missing Determiner (62%) Missing Preposition (18%) Missing Conjunction (13%) Missing Verb (4%) Missing Subject (3%)
Insertion (17%)	Extra Preposition (36%) Extra Determiner (26%) Extra Conjunction (20%) Extra Verb (15%) Extra Intensifier (3%)

substitution, insertion, and deletion. For each category, we listed the five most common error types (Table 3) which account for 73% of the errors. As Foster (2007; 2005) and Lee (2009) generated a treebank of ungrammatical English, we also produced artificial grammar errors systemically. The error generation procedure takes as input a part-of-speech tagged sentence which is assumed to be well-formed, and outputs a part-of-speech tagged ungrammatical sentence. In the first step of the error generation procedure, we set the Grammar Error Rate (GER) between 0 % ~ 100 % and determined error counts to be produced based on the GER. Then, we distributed the errors among categories and error types according to the percentages in the error types list (Table 3).

5.3 Results

5.3.1 Hybrid model vs. Utterance model

To verify the effectiveness of the dialog state-awareness, we compared the hybrid model with the utterance only model. The utterance only model just omits the dialog context model from the hybrid model. We conducted 200 dialogs for each model per 10 % GER intervals. The hybrid model significantly outperformed the utterance only model for overall range of GER. As the GER increased, the performance of the utterance only model decreased dramatically, whereas the performance of the hybrid model decreased smoothly (Fig. 3). It verifies the effectiveness of dialog state-awareness through our hybrid approach.

5.3.2 Appropriateness of feedback

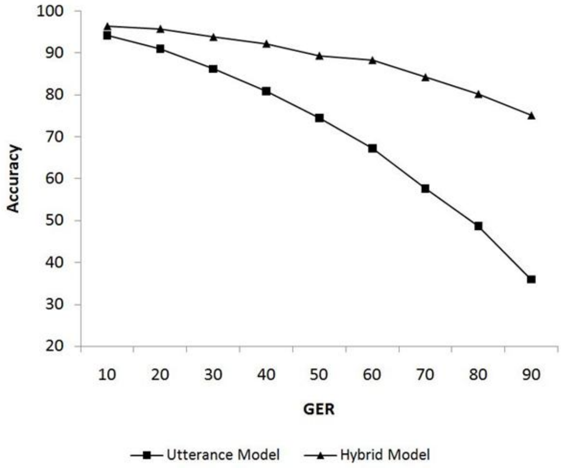

Figure 3: Comparison between the hybrid model and the utterance only model

On the contrary to the task oriented dialogs, language tutoring systems do not need to exactly recognize the learner's intention. Even if the inferred intention is not the same as the actual one, the feedback can be valuable for language acquisition as long as the feedback is appropriate to the dialog context. In fact, human tutors also generate feedback relying on only the dialog context when the learners' utterances are highly incomprehensible. Often, without such feedback, the conversation even can be stuck with a learner's problematic utterance, thereby cannot successfully finish. As McClelland (1961) noted the role of success motivation in learning, the completion of a dialog itself is undoubtedly important rewards in foreign language learning. Therefore, we want our hybrid model to provide feedback appropriate to the dialog context regardless

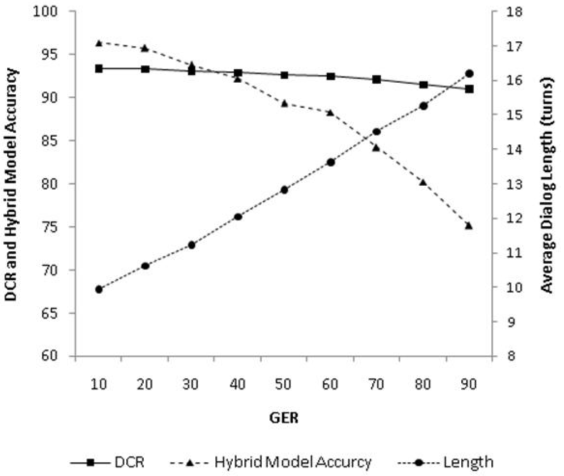

Figure 4: The relation between Dialog Completion Rate and the performance of the hybrid model and the Average Dialog Length

of whether the inferred intention exactly equals the original one or not.

To evaluate the appropriateness of the feedback, we conducted 200 dialogs per 10 % GER intervals from 10 % to 90 %, and observed the Dialog Completion Rate (DCR) as the GER increased. As the GER increased, the performance of the hybrid model decreased, whereas the DCR decreased very slightly (Fig. 4). Because of the clarification sub-dialogs, the average dialog length increased as the GER increased. Based on this result, we can conclude that our method is suitable to produce appropriate feedback even when the inferred intention is not the same as the actual one. This is because the dialog context model effectively confines candidate intentions within the given context.

6 CONCLUSION

When language learners speak incomprehensible utterances, a DB-CALL system should provide matching fluent utterances. We proposed a novel hybrid model that allows natural decomposition between the utterance model and the dialog context model in terms of the psychological language production process. It led to an efficient way for adjusting to diverse fluency levels with minimal efforts. In addition, our elaborate dialog context model using enhanced k-nearest neighbors algorithm gave rise to more accurate inference of the language learners' intention. Also, it proved to be effective to provide appropriate context-aware feedback so that the learners can obtain positive rewards by successfully completing dialogs.

To further confirm the pedagogic value of the method, we plan to design the pretest-posttest comparison of an experimental and a control group. In addition, we want to simulate ASR errors of non-native speakers so that we can test whether our method is also effective in a speech environment.

ACKNOWLEDGEMENTS

This research was supported by the Web-Scale NLP Theme Program funded by Microsoft Research Asia

REFERENCES

Ai, H., Roque, A., Leuski, A., Traum, D. 2007. Using Information State to Improve Dialog Move Identification in a Spoken Dialog System In Proc. Interspeech 2007, Antwerp

Bohus, D. and Rudnicky, A. 2003. RavenClaw: Dialog Management Using Hierarchical Task Decomposition and an Expectation Agenda. In Proc. Eurospeech 2003, Geneva.

Carroll, D. 2003. The Psychology of Language Wadsworth Publishing Co Inc. 4th edition

Dasarathy, B. V. 1990. Nearest neighbor (NN) norms: NN pattern classification techniques Los Alamitos: IEEE Computer Society Press

Foster J. 2007. Treebanks Gone Bad: Parser evaluation and retraining using a treebank of ungrammatical sentences. International Journal on Document Analysis and Recognition, 10(3-4), 129--145.

Foster J. 2005. Good Reasons for Noting Bad Grammar: Empirical Investigations into the Parsing of Ungrammatical Written English. PhD Dissertation, Department of Computer Science, Trinity College, University of Dublin

Jung S., Lee C., Kim K., Lee G. G. 2009. Data-driven user simulation for automated evaluation of spoken dialog systems. Computer Speech and Language. 23(4):479-509, Oct 2009

Lee C., Jung S., Kim S., Lee G. G. 2009. Example-based Dialog Modeling for Practical Multi-domain Dialog System. Speech Communication. 51(5):466-484, May 2009.

Lee S., Lee G. G. 2009. Realistic grammar error simulation using markov logic. To be published in Proceedings of the ACL 2009, Singapore, August 2009

Long, M. H. 2005. The role of the linguistic environment in second language acquisition in W. C. Ritchie and T. K. Bhatia (eds): Handbook of Second Language Acquisition. New York: Academic Press, pp. 413–68.

McClelland, D. C. 1961. The achieving society New York: Van Nostrand

Michaud L. N. 2002. Modeling User Interlanguage in a Second Language Tutoring System for Deaf Users of American Sign Language. PhD Dissertation, Department of Computer and Information Sciences, University of Delaware

Morton, H. and Jack, M. A. 2005. Scenario-Based Spoken Interaction with Virtual Agents, Computer Assisted Language Learning, 18:3 pp. 171-191.

Poesio, M. and Mikheev, A. 1998. The Predictive Power of Game Structure in Dialog Act Recognition: Experimental Results Using Maximum Entropy Estimation. In Proceedings of ICSLP-98.

Ratnaparkhi, A. 1998. Maximum Entropy Models for Natural Language Ambiguity Resolution, PhD thesis, University of Pennsylvania, 1998.

Raux, A. and Eskenazi, M. 2004. Using Task-Oriented Spoken Dialog Systems for Language Learning. Potential, Practical Applications and Challenges, Proceedings INSTIL.

Schneider, D. and McCoy K. F. 1998. Recognizing Syntactic Errors in the Writing of Second Language Learners. In Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and the Seventeenth International Conference on Computational Linguistics (COLING-ACL), Volume 2

Swain, M. 1996. The output hypothesis: Theory and research in E. Hinkel (ed.): Handbook of Research in Second Language Teaching and Learning. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 471–83.

.