Affective-Cognitive Dialogue Act Detection in an Error-Aware Spoken Dialogue System

Wei-Bin Liang, Chung-Hsien Wu, and Meng-Hsiu Sheng
Department of Computer Science and Information Engineering,National Cheng Kung University, Tainan, Taiwan
E-mail: {liangnet, chunghsienwu}@gmail.com

***Abstract*—This paper presents an approach to affective-cognitive dialogue act detection in a spoken dialogue. To achieve this goal, the input utterance is decoded as the affective state by an emotion recognizer and a word sequence by an imperfect speech recognizer separately. Besides, four types of evidences are employed to grade the score of each recognized word. The recognized word sequence is used to derive the candidate sentences to alleviate the problem of unexpected language usage for the cognitive state predicted by the vector space-based dialogue act detection. The Boltzmann selection based method is then employed to predict the next possible act in the spoken dialogue system according to the affective-cognitive states. A model of affective anticipatory reward that is assumed to arise from the emotional seeking system is adopted for enhancing the efficacy of dialogue act detection. Finally, the evaluation data are collected and the experimental results confirm the improved performance of the proposed approach compared to the baseline system on the task completion rate.**

TABLE I
EXAMPLE OF ERROR-AWARE INTERACTION IN A DIALOGUE. *S* INDICATES A SYSTEM TURN AND *U* INDICATES A USER TURN.

No	Spkr.	Turn
1	S:	What can I help you with?
2	U:	McDonald's promotion [**Michael Bolton**]
3	S:	Sorry, I don't understand "Michael Bolton".
4	U:	Is there any McDonald's promotion?
5	S:	Wait, I am checking the list of promotion.
6	S:	Hmmm, buy one McFlurry, get one free.
7	U:	It is lunch time.
8	S:	Most of McDonald's is open 24 hours a day in Taiwan.
9	U:	I know but I am so hungry.
10	S:	You can have many McFlurry to eat.
11	U:	McFlurry is a dessert not meal!
12	S:	Oops! You want to know McDonald's value meal. Is it right?
13	U:	Yes!

I. INTRODUCTION

Spontaneous speech is a fundamental and natural means of human-machine communication. Services based on the spoken dialogue systems (SDSs) have been deployed in a wide range of goal-orientated applications, for example, airline travel information system (ATIS) for flight information [1] and AT&T "How May I Help You" for call routing [2]. However, understanding what users like to do/need to get is critical in human computer communication. To realize this kind of application system, the dialogue act (DA) is used to describe dialogue behavior that a DA represents the meaning of an utterance at the level of illocutionary force [3][4][5][6][7][8]. One of the greatest challenges when building SDSs is to deal with uncertainty and errors [9]. Uncertainty comes partly from the ambiguity of natural language itself. Moreover, a great deal of uncertainty comes from imperfect automatic speech recognition (ASR). The errors resulting from error-prone recognition of spontaneous speech with disfluencies [10][11] and pronunciation variations [12] leads to output uncertain DAs which will dramatically degrade the performance for language understanding and miscommunication between humans and computers.

Herein, the term *Error-Aware* is defined as the ability to recover the user's intents from misunderstanding. For a spoken dialogue system, the misunderstanding can be made by misrecognition from an imperfect automatic speech recognition (ASR) component or uncertain DA from the spoken language understanding (SLU) component in an SDS. Table I illustrates an example of error-aware in a dialogue. Even the ASR can perfectly transcribe the user's utterances, the system may still misunderstand the user's intent. In Turn 2, the user replies with his/her query which is then misrecognized as saying "'Michael Bolton". Hence, the user start the interaction over in Turn 4. In Turn 5, the user's actual dialogue act is a query about meal but it is misrecognized as a query about promotion. Thus, the user's reaction was expressed in an unhappy emotion in Turn 11 so that the system is aware of the real user's intent.

This interaction illustrates the motivation for the three main approaches that have been developed in order to minimize the effects of errors and uncertainty in the dialogue management (DM) module of an SDS [13][14][15]. First, the most commonly used measure to identify the errors locally is the confidence score [16]. The decision engaged in this method is typically to compare the confidence score of a recognition hypothesis against the manually predefined threshold. However, confidence scores are not entirely reliable and are dependent on noisy environments and user types. Moreover, false acceptance may not be easy for the user to correct the system and put the dialogue back on track. Second, accepting those misrecognitions is inevitable so that their consequences are difficult to anticipate. To address the problem of false acceptance, the DM can adopt some error recovery strategies [9][17]. As Clark points out, when humans speaks to each other, there is a collaborative process of avoiding and recovering from miscommunication that often goes unnoticed by providing and evaluating positive and negative evidence of understanding – a process commonly referred to as grounding [18]. Instead of generating a clarification subdialogue,

Skantze chose a specific interpretation and run the risk of making a mistake as the grounding decision problem [9] and then employed hand-crafted confidence thresholds for different levels of evidence of understanding, for example, acceptance, implicit/explicit confirmation and rejection [19]. But these hand-crafted confidence thresholds are often difficult to determine. Finally, accepting those misrecognitions will be passed to the dialogue state maintained by the system; thus, it seems unwise to maintain just one hypothesis for the current dialogue state. Recently, two major stochastic dialogue modeling approaches are developed for the DM component. The first one approach using reinforcement learning based on Markov decision process (MDP) [20] or partially observable MDP (POMDP) [14][21][22] is widely used to determine which action a system should take in a given situation. Another one is to employ the finite-state transducer (FST)-based techniques to automatically create the DM component from an n-gram model of a tag sequence [23][24]. Although the stochastic approaches can be done automatically and requires little human supervision, automated planning is hard due to unexpected language usage. Moreover, affective computing is not concerned in the current techniques of DM modeling.

In this paper, we propose a statistical grounding decision process to automatically determine the confidence thresholds for further accepting inevitable misrecognitions. Moreover, the partial expansion tree is also proposed to decrease the effect of misrecognitions and unexpected language usage. Unlike the conventional DAs represented by the semantic frame, we develop a FST-like technique to form the DA. Thus, the vector space model-based approach is employed to construct a DA detection model using various linguistic information. Finally, the softmax decision is utilized to construct the affective-cognitive DM in order to take into account the emotion and DA.

The rest of this paper is concerned with the design issue of affective-cognitive dialogue act detection and the ability of error-awareness. We start with a overall illustration of system framework in Section II. From Section IV to Section VI, we will introduce those components for decoding the input utterance, statistical grounding decision process, candidate sentence generation, and dialogue act detection. We also give a detailed illustration of the affective-cognitive DM in Section VII Finally, we conclude with the evaluation results in Section VIII.

II. Framework

In this paper, the error-aware DA detection is illustrated by a typical block diagram of an SDS shown in Fig. 1. An input utterance is decoded into a word sequence $\mathbf{W}$ from an ASR component and an emotional state E_t^* from a support vector machine (SVM)-based emotion recognition component, respectively. In the grounding component, each recognized word in $\mathbf{W}$ is assigned an understanding evidence (UE) by the z-score based confidence measure (CM) for grounding decision process, instead of rejecting unreliable recognized words directly. Then, because the confidence measures are not entirely reliable and the user's language usage is often unexpected, a partial expansion tree (PET) component is utilized to derive the word sequence $\mathbf{W}$ to several candidate utterances. The DA detection will be done in the SLU component. Due to different language usages for a query, each predefined DA type is partitioned into several subtypes by the k-means clustering algorithm using the linguistic features of PET, including word n-grams and syntactic rules obtained from the Stanford parser [25]. The DA detection model is to model the relation between each subtype of a DA and all linguistic features based on the technique of latent semantic analysis (LSA). Therefore, a cosine measure can be easily used to detect a DA. The DM component is employed to make a decision of current interaction state using the most likely emotional state E_t^* and optimal DA A_t^*. Moreover, the affective anticipation estimate computes the decision probability conditioned on the current affective state and cognition state based on the Boltzmann selection. At last, the DM outputs a response to the speaker.

III. Speech Decoding

For the goal of affective-cognitive dialogue act detection, firstly, the user' input utterance U_t will be decoded by two different recognizers. One is the ASR component implemented by the hidden Markov model toolkit (HTK) [26] and is employed to transcribe the utterance U_t to a recognized word sequence $\mathbf{W}$. Another is the emotion recognition component. The SVM is a widely popular emotion recognizer [27] and can be easily implemented by a library for support vector machines (LIBSVM) [28]. Thus, the most likely emotional state can be decoded and defined as:

$$E^* = \max_{E_i \in \Omega_E} Pr(U_t|E_i) \tag{1}$$

IV. Grounding

In [18], *grounding* is defined to establish a thing as part of common ground well enough for current purpose. Thus, the requirements on how many evidences are needed vary depending on the current purpose. For each recognized word w in $\mathbf{W}$, the z-score is employed to assign the UE and defined as

$$z(w) = \frac{f(w) - \mu(w)}{\sigma(w)} \tag{2}$$

where $f(w)$ is the score of w; $\mu(w)$ and $\sigma(w)$ are the mean and standard deviation of the score of w, respectively. By means of z-score, the confidence thresholds of the grounding decision process can be determined statistically. So, the UE of w is defined as

$$UE(w) = \begin{cases} \text{Accept} & \text{if} \quad z(w) > \theta_3 \\ \text{Uncertain Accept} & \text{if} \quad \theta_2 < z(w) \leq \theta_3 \\ \text{Uncertain Reject} & \text{if} \quad \theta_1 < z(w) \leq \theta_2 \\ \text{Reject} & \text{if} \quad z(w) < \theta_1 \end{cases} \tag{3}$$

where θ_1, θ_2, and θ_3 are determined by two normal distribution of ASR output shown in Fig. 2.

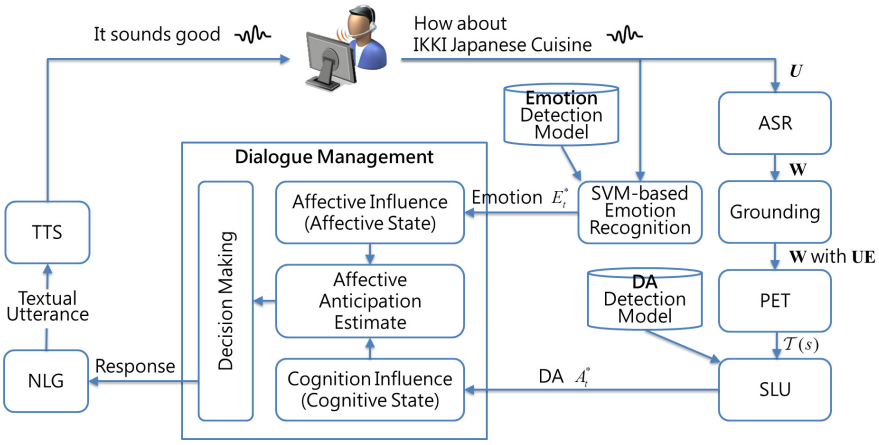

Fig. 1. Block diagram of the DA detection with error-awareness in a spoken dialogue system

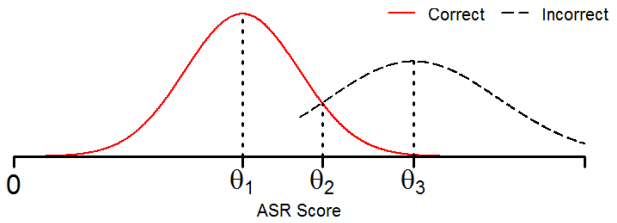

Fig. 2. Illustration of the thresholds of different evidences

V. Partial Expansion Tree

Although the UE aids to develop error recovery strategies, the user's unexpected language usage usually confuses the system. To alleviate such problem, the partial expansion tree (PET) is proposed to derive several candidate utterances. In an SDS, it is often beneficial to define a set of keywords $\mathcal{K}$ and a set of non-keywords $\mathcal{N}$. Each word $w \in \mathcal{K}$ should be indicative of the DA of the sentence. The set of sentences $\mathcal{S}$ containing at least one keyword in $\mathcal{K}$, can be represented as $\mathcal{S} = \mathcal{N}^* (\mathcal{K} \mathcal{N}^*)^+$, where $\mathcal{K}^+$ means a string of one or more words in $\mathcal{K}$. Given a sentence $s \in \mathcal{S}$, a partial sentence is formed by keeping all the keywords in s and some of the non-keywords in s. These partial sentences can be compiled in a tree, called the partial expansion tree and denoted as $\mathcal{T}(s)$.

VI. Spoken Language Understanding

In this paper, DA detection is employed to realize the spoken language understanding component and can be formulated as follows. At turn t, the most likely DA is determined by

$$\begin{aligned} A_t^* &= \arg\max_{A\in\Omega} \max_{\mathbf{W}} Pr(A, \mathbf{W}|U_t) \\ &= \arg\max_{A\in\Omega,\mathbf{W}} Pr(\mathbf{W}|U_t) Pr(A|\mathbf{W}, U_t), \end{aligned} \tag{4}$$

where U_t is the user's input utterance, $\Omega = \{A_1, \ldots, A_q\}$ is the set of DAs, and $\mathbf{W}$ is the most likely ASR output. The ASR-related first term in Eq.(4) is introduced as the ASR score function and defined as

$$Pr(\mathbf{W}|U_t) \propto f(\mathbf{W}, U_t). \tag{5}$$

In addition, assuming that the information provided by U_t is completely conveyed in $\mathbf{W}$, we can approximate the second term in Eq.(4) by one function

$$Pr(A|\mathbf{W}) \propto g(A, \mathbf{W}), \tag{6}$$

where $g(A, \mathbf{W})$ is introduced as the lexical score function of DA detection. Thus, Eq.(4) can be re-written as

$$A_t^* \approx \arg\max_{A\in\Omega,\mathbf{W}} f(\mathbf{W}, U_t)\ g(A, \mathbf{W}). \tag{7}$$

We will specify and explain how the score in Eq.(7) are computed.

A. Feature Extraction

The DAs are conventionally represented by the semantic frame; however, such bag-of-slots based representation should be designed carefully. So, the finite-state transducer is employed to represent a DA in our system. Moreover, the proposed representation must be employed to generate the DA. Hence, the semantic class will be also included in feature extraction. To reach this purpose, two major categories of features are extracted from the manually transcribed sentences and the results of PET. The first category is the syntactic rules obtained from a probabilistic context free grammar parser (Stanford Parser [25]) trained on the annotated parts-of-speech (POSs). Because the linguistic property of Chinese, the POSs are treated as the basic feature of the vector space in the detection model. For example, the word "respect" is a noun but also a verb in Chinese and English, so it is distinguishable by the POS. The second one is the n-gram of named entities and semantic classes shown in Fig. 3. Herein, the n-gram related features are considered as the certainly influential features because they are somewhat used to represent the weighted frames of tokens. By these features, an M dimensional feature vector for each sentence can be extracted for model training. In the training phase, the manually transcribed sentences and

Class	Named Entities
RESTAURANT	IKKI, TASTY, McDonald's, Pizza-Hot, KFC, BurgerKing
TIME	Morning, Afternoon, Evening, Night
MEAL	Burger, Sushi, Steak, Pork

Fig. 3. Example of semantic class

the results of PET are included to increase the frequencies of named entities and semantic classes.

B. DA Clustering

At the beginning of system construction, types of DA and their corresponding slots are not easy to design perfectly. In addition, due to the user's unexpected language usage, a DA can be queried by varied types of sentences. For example, a query sentence can begin with "Can you tell me" or "I want to know". So, a mechanism is needed to generate the DA types. Herein, the k-means clustering algorithm is employed to partition a i-th base dialogue act A_i into k sub-DAs, i.e. $A_i = \{a_{i,1}, \ldots, a_{i,k_i}\}$. For sentence clustering, a sentence S_i is represented as $S_i \equiv (\delta_{i,1}, \delta_{i,2}, \ldots, \delta_{i,M})$ where $\delta_{i,m}$ equals 1 if S_i includes the m-th feature; otherwise it is set to 0. Furthermore, the similarity measure between two sentences can be estimated using the cosine distance measure defined as

$$\text{Similarity}(S_j, S_k) = \frac{S_j \cdot S_k}{|S_j| \cdot |S_k|} \tag{8}$$

C. DA Modeling

According to previous procedures, the feature-by-DA matrix Φ is constructed as

$$\Phi = \begin{array}{c} \\ \delta_1 \\ \vdots \\ \delta_M \end{array} \begin{array}{c} \begin{array}{cccccccc} a_{1,1} & \cdots & a_{1,k_1} & \cdots & a_{q,1} & \cdots & a_{q,k_q} \end{array} \\ \begin{pmatrix} \phi^1_{1,1} & \cdots & \phi^1_{1,k_1} & \cdots & \phi^1_{q,1} & \cdots & \phi^1_{q,k_q} \\ \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ \phi^M_{1,1} & \cdots & \phi^M_{1,k_1} & \cdots & \phi^M_{q,1} & \cdots & \phi^M_{q,k_q} \end{pmatrix} \end{array} \tag{9}$$

Then, the LSA-based technique with entropy-based weighting scheme [29] is employed to model the importance between features and sub-DAs.

D. Score of DA detection

In our system, the lexical score $g(A, \mathbf{W})$ in (7) is further broken into two terms

$$g(A, \mathbf{W}) \approx g_F(s, a) g_A(a, A) \tag{10}$$

where $g_F(\mathbf{f}, a)$ is called the feature-level score and $g_A(a, A)$ is called the prior score. Note that s denotes the sentence after text processing. The cosine distance measure is employed for the feature-level score estimation,

$$g_R(s, a = a_j) = \max_{\sigma \in \mathcal{T}(s)} \frac{\mathbf{b}_\sigma^T \mathbf{a}_j}{|\mathbf{b}_\sigma||\mathbf{a}_j|} \tag{11}$$

where $\mathbf{b}_\sigma^T$ is the vector representation (using the coordinates of the features) of a candidate sentence σ in $\mathcal{T}(s)$, and $\mathbf{a}_j$ is the j^{th} column vector in the matrix Φ. For the prior score, we use the approximation

$$g_A(a = a_j, A) = \frac{Pr(a_j)}{\sum_q Pr(a_q)} \tag{12}$$

In other words, the prior score is the occurrence probability of sub-DA a_j conditioned on the base DA A and can be estimated from a training corpus by relative frequencies.

VII. Decision-Making

The DM component should respond to the user conditioned on the relationship between current affective states and cognitive states, that is, the response should be selected by fusion of these two current states. However, it is difficult to extract and learn such relationship through interaction with the users. Reinforcement learning (RL) methods build agents that maximize their expected utility only using reward from the environment [30]. Accordingly, the RL does not use the explicit teaching signals, but uses evaluative feedback obtained through the interaction with the environment. Moreover, Q-learning is one of the most important breakthroughs in RL and is an off-policy temporal-difference control algorithm [31]. Its simplest method, one-step Q-learning, for assigning value to state-action pair based on incremental dynamic programming learns the following action-value function

$$\begin{aligned} Q(s_t, a_t) \leftarrow & Q(s_t, a_t) \\ & + \alpha \left[R_{t+1} + \eta \max_{a \in A} Q(s_{t+1}, a) Q(s_t, a_t) \right] \end{aligned} \tag{13}$$

where R_{t+1} is the reward value, α is a learning rate ($0 < \alpha < 1.0$), η is a temporal discounting rate ($0 < \eta < 1.0$), and A is a set of action, e.g., confirmation and start over. Additionally, it is hard to design the reward function for R because the reward value should be extracted through the interaction with the SDS. Herein, we assume that the user's emotional states reflect the correctness of a system's responses. Therefore, we use the following reward function

$$R_t \leftarrow \gamma(a_p) \cdot R_0 \tag{14}$$

where $\gamma(\cdot)$ is a Sigmoid function, a_p is the detected DA, R_0 is a base value for the reward. Therefore, R_t will be low as the user is not satisfied with the system's responses. For the value function $Q(s, a)$, it is simply defined as

$$Q(s, a) \stackrel{def}{=} Pr(A, \mathbf{W}|U_t) + Pr(U_t|E) \tag{15}$$

The decision-making model $Pr(d|E, a)$ computes the probability of choosing a decision d conditioned on both of the current cognitive state a and the affective state E, and uses the Boltzmann selection that can easily control the trade-off between exploration and exploitation through the inverse temperature β In this paper, we apply the softmax action selection (Boltzmann selection), and the i-th decision d_i is selected with following probability,

$$Pr_t(d_i) = \frac{exp(\beta \cdot Q(E_t, a_t, d_t))}{\sum_{d=1}^{|D|} exp(\beta \cdot Q(E_t, a_t, d))} \tag{16}$$

where β is inverse temperature. When β is low, the DM randomly selects a reaction. As the β increase, the DM deterministically selects the reaction with high decision strength. After making decision, the SDS can select a reaction to respond to the user according to the current cognitive and affective state.

VIII. Evaluation

A. Corpus Collection

For corpus collection, we assume that the system will not understand the user's goal if the operator has no idea about the textual utterance. Therefore, a modified Wizard-of-Oz method was employed to collect the corpus about the off-campus restaurant information. The operator supervised the textual utterances obtained from the ASR and even changed the system act to the user. By means of this approach, mixed-initiative strategy is adopted to prevent from serious misunderstanding. Additionally, the restaurant information was obtained from the internet, such as blogs, Google Map, and bulletin board systems. The collected data contain query words for the system functions, including system service, restaurant information, food information, and greeting/ending. From the corpus, a total of $1,697$ utterances of 38 base DAs were collected in 118 dialogues wherein 383 utterances of 28 DAs are annotated as misunderstanding. This collected corpus comprises 432 tokens and are used to derive 682 syntactic rules. To avoid incorrect parsing and clustering, 26 semantic classes (e.g., RESTAURANT) were manually defined to replace the named entities (e.g., PizzaHot). Moreover, each utterance corresponds to a DA. The HTK-based ASR trained on the TCC-300 corpus and adapted by the collected data was constructed to achieve 78% recognition accuracy.

B. Emotion Recognition

To recognize the user's emotional states, each collected utterance is also simply labelled as "Positive" or "Negative" according to the manner of speaking. As the example shown in Table I, the user was not satisfied with the system response in Turn 10 so the user answered the system angrily in Turn 11. Thus, Turn 11 should be annotated as "Negative". Totally, 361 out of 383 misunderstanding utterances are labelled as negative utterance. To train the SVM-based emotion recognizer, the scripting capabilities of the freely-available Praat [32] software is emolpyed to process the collected corpus and extract the prsodic feature set composed pitch-related, intensity-related, formant-related features and their functionals (e.g., mean, standard deviation, maximum and minimum). Finally, 68 dimension feature vector is employed to train the emotion recognition model. In this paper, the SVM-based emotion recognizer achieved 81.46% accuracy.

C. Dialogue Act Detection

Because the collected data is not sufficiently extensive to be divided into training, development, and test datasets, we used a five-fold cross-validation method to evaluate the proposed approach. Table II shows the average detection rate of the base DA (row captioned as Base-DA) and misunderstanding DA (row captioned as Mis-DA) for each method combination. The KW set is that the task is conducted based on keyword spotting. The SR set, ngram set, and SC set were designed to incrementally assess the effects of syntactic rule, n-gram of keyword, and n-gram of semantic class on the detection models, respectively. One can observe that these three types of features can contribute to our system, especially the semantic class. The n-gram related features confirmed that the weighted frame of tokens can alleviate the fake keywords. By contrast, our system outperformed the keyword-based approach because of the linguistic property of Chinese. In this paper, we concern the sensitivity of imperfect ASR so the Trans set was conducted by the manually transcribed utterances, i.e., the result is the upperbound of our system. Moreover, the detection accuracy of misunderstanding DA is 66%.

TABLE II
Average detection rate (%) of each method combination

	KW	SR	$n$gram	SC	Trans
Base-DA	50.16	72.10	75.39	83.95	94.57
Mis-DA	51.96	78.85	80.68	81.46	92.69

D. Dialogue Completion Rate

Here, we will evaluate the performance of Boltzmann selection-based dialogue management conditioned on the user's current affective and cognitive states. In order to have a running estimate of success rate, we built a prototype for the query system to assess the dialogue completion rate. Finally, the SDS with affective-cognitive decision making results in an average task success rate of 92.4%.

IX. Conclusion

This paper investigated the topic of affective-cognitive dialogue act detection based on error awareness for spoken dialogue error handling. For the incompletely reliable confidence measure, z-score based statistical grounding decision process is proposed to grade each recognized word obtained from an imperfect ASR. For the unexpected language usage, the partial expansion tree is employed to derivate candidate sentence for further DA detection. LSA-based DA detection model that considers error awareness is beneficial for DA detection. Error handling actions were integrated into a Boltzmann selection-based dialogue management for decision-making according to the affective and cognitive states. In addition, predefined response types were employed for different repair conditions. The experiments indicate that the average detection accuracy of the LSA-based approach is 82.7%, and the dialogue management with error handling achieves an average task success rate of 92.4%. These results indicate that the proposed SDS not only alleviates the degraded performance of DA detection because of the error-prone ASR, but also improves the task success rate.

REFERENCES

[1] P. J. Price, "Evaluation of spoken language systems: the atis domain," in *Proc. the workshop on Speech and Natural Language*, 1990.

[2] A. Gorin, G. Riccardi, and J. H. Wright, "How may I help you?" *Speech Communication*, vol. 23, pp. 113–127, 1997.

[3] J. L. Austin, *How to Do Things With Words*. Clarendon Press, Oxford, 1962.

[4] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *COMPUTATIONAL LINGUISTICS*, vol. 26, pp. 339–373, 2000.

[5] C.-H. Wu and G.-L. Yan, "Speech act modeling and verification of spontaneous speech with disfluency in a spoken dialogue system," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 330–344, May 2005.

[6] N. Webb, "Cue-based dialogue act classification," Ph.D. dissertation, University of Sheffield, 2010.

[7] W.-B. Liang, C.-H. Wu, and Y.-C. Hsiao, "Dialogue act detection in error-prone spoken dialogue systems using partial sentence tree and latent dialogue act matrix," in *Proc. INTERSPEECH'2010*, Sep 2010, pp. 3038–3041.

[8] C.-P. Chen, C.-H. Wu, and W.-B. Liang, "Robust dialogue act detection based on partial sentence tree, derivation rule, and spectral clustering algorithm," *EURASIP J. Audio, Speech and Music Processing*, vol. 2012, p. 13, 2012.

[9] G. Skantze, "Error handling in spoken dialogue systems - managing uncertainty, grounding and miscommunication," Ph.D. dissertation, Speech Communication KTH, Stockholm, Sweden, Nov 2007.

[10] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1526–1540, Sep 2006.

[11] J.-F. Yeh and C.-H. Wu, "Edit disfluency detection and correction using a cleanup language model and an alignment model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1574–1583, Sep 2006.

[12] W.-B. Liang, C.-H. Wu, and Y.-K. Kang, "Recognition of syllable-contracted words in spontaneous speech using word expansion and duration information," in *Proc. International Symposium on Chinese Spoken Language Processing (ISCSLP'2008)*, Dec 2008, pp. 1–4.

[13] M. McTear, I. O'Neill, P. Hanna, and X. Liu, "Handling errors and determining confirmation strategies - an object-based approach," *Speech Communication*, vol. 45, no. 3, pp. 249–269, 2005, ¡ce:title¿Special Issue on Error Handling in Spoken Dialogue Systems¡/ce:title¿.

[14] J. Williams and S. Young, "Scaling up pomdps for dialog management: The "summary pomdp" method," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'2005)*, Nov 2005, pp. 177–182.

[15] C. Lee, S. Jung, K. Kim, D. Lee, and G. G. Lee, "Recent approaches to dialog management for spoken dialog systems," *Journal of Computing Science and Engineering*, vol. 4, no. 1, pp. 1–22, 2010.

[16] W. K. Lo and F. Soong, "Generalized posterior probability for minimum error verification of recognized sentences," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '2005)*, vol. 1, 2005, pp. 85–88.

[17] A. Raux, D. Bohus, B. Langner, A. W. Black, and M. Eskenazi, "Doing research on a deployed spoken dialogue system: One year of let¡s go! experience," in *Proc. INTERSPEECH'2006*, 2006, pp. 65–68.

[18] H. H. Clark, *Using Language*. Cambridge University Press, 1996.

[19] G. Bouwman, J. Sturm, and L. Boves, "Incorporating confidence measures in the dutch train timetable information system developed in the arise project," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'1999)*, vol. 1, 1999, pp. 493–496.

[20] E. Levin, R. Pieraccini, and W. Eckert, "A stochastic model of human-machine interaction for learning dialog strategies," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 1, pp. 11–23, 2000.

[21] D. Kim, J. Kim, and K.-E. Kim, "Robust performance evaluation of pomdp-based dialogue systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 1029–1040, May 2011.

[22] S. Varges, G. Riccardi, S. Quarteroni, and A. Ivanov, "Pomdp concept policies and task structures for hybrid dialog management," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2011)*, May 2011, pp. 5592 –5595.

[23] C. Hori, K. Ohtake, T. Misu, H. Kashioka, and S. Nakamura, "Weighted finite state transducer based statistical dialog management," in *Proc. IEEE Workshop on Automatic Speech Recognition Understanding (ASRU'2009)*, 2009, pp. 490–495.

[24] ——, "Statistical dialog management applied to wfst-based dialog systems," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2009)*, 2009, pp. 4793–4796.

[25] R. Levy and C. D. Manning, "Is it harder to parse chinese, or the chinese treebank?" in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL'2003)*, vol. 1, 2003, pp. 439–446.

[26] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[27] C.-H. Wu and W.-B. Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels," *IEEE Transactions on Affective Computing*, vol. 2, no. 1, pp. 10–21, 2011.

[28] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 1–27, 2011.

[29] J. Bellegarda, "Exploiting latent semantic information in statistical language modeling," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1279–1296, Aug 2000.

[30] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[31] C. Watkins, "Learning from delayed rewards," Ph.D. dissertation, Cambridge, May 1989.

[32] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," [Computer program], Jun 2013, ver. 5.3.51, retrieved 2.