Wu et al. EURASIP Journal on Audio, Speech, and Music Processing (2017) 2017:9
DOI 10.1186/s13636-017-0107-3

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Miscommunication handling in spoken dialog systems based on error-aware dialog state detection

Chung-Hsien Wu*, Ming-Hsiang Su and Wei-Bin Liang

Abstract

With the exponential growth in computing power and progress in speech recognition technology, spoken dialog systems (SDSs) with which a user interacts through natural speech has been widely used in human-computer interaction. However, error-prone automatic speech recognition (ASR) results usually lead to inappropriate semantic interpretation so that miscommunication happens easily. This paper presents an approach to error-aware dialog state (*DS*) detection for robust miscommunication handling in an SDS. Non-understanding (*Non-U*) and misunderstanding (*Mis-U*) are considered for miscommunication handling in this study. First, understanding evidence (UE), derived from the recognition confidence, is adopted for Non-U detection followed by Non-U recovery. For Mis-U with the recognized sentence containing uncertain recognized words, the partial sentences obtained by removing potentially misrecognized words from the input utterance are organized, based on regular expressions, as a tree structure to tolerate the deletion or rejection of keywords resulting from misrecognition for Mis-U DS modeling. Latent semantic analysis is then employed to consider the verified words and their *n*-grams for DS detection, including Mis-U and predefined *Base* DSs. Historical information-based *n*-grams are employed to find the most likely DS for the SDS. Several experiments were performed with a dialog corpus for the restaurant reservation task. The experimental results show that the proposed approach achieved a promising performance for Non-U recovery and Mis-U repair as well as a satisfactory task success rate for the dialogs using the proposed method.

Keywords: Error-aware dialog act, Miscommunication, Spoken dialog systems

1 Introduction

In recent years, voice-driven human-computer interaction has benefited greatly from steady improvements in the underlying speech technologies, such as speech recognition, speech synthesis, natural language understanding, and machine learning [1]. Spoken dialog systems (SDSs) are supposed to enable an efficient and intuitive communication between humans and computers [2], and help users achieve the goal which they want to accomplish by using spoken languages. This could be done by mapping the spoken utterance to the semantic meaning of the recognized word sequence using the automatic speech recognition (ASR) technology. Then spoken language understanding (SLU) maps the semantic meaning of the recognized word string to the user's semantic slots and a list of values [3]. The semantic slots and their corresponding values are maintained by a dialog state (DS) tracking component over dialog turns, and change gradually over the process of the dialog.

Dialog state tracking (DST) is one of the key sub-tasks of dialog management, which updates the dialog states at each moment on a dialog conversation [4]. Because the errors from ASR and SLU are generally encountered, DST task facing the errors may lead to misunderstanding of the user's intention. In previous studies, several methods have been proposed to deal with the problems on ASR and SLU errors for performance improvement. Some typical examples of the proposed approaches include handcrafted rule-based methods [5, 6], Bayesian networks [7, 8], discriminative models [9], and long short-term memory neural networks [10]. As the dialog system, one of the prominent human-computer interaction research areas, has been applied to a wide range of domains from simple

* Correspondence: chunghsienwu@gmail.com
Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan

Springer Open

© The Author(s). 2017 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

goal-oriented applications to complex conversational systems, a common testbed or evaluation measure for this task is highly desirable. To provide a common testbed for the DST task, a series of the Dialog State Tracking Challenge (DSTC1~4) has been successfully organized and completed in the past years. This challenge task series has spurred significant work on dialog state tracking, yielding both numerous new techniques as well as a standard set of evaluation metrics [4].

At the early stages, SDSs often used handcrafted rules for DS tracking with corresponding confidence scores [5, 11]. Such SDSs did not require data to implement and provide an accessible method to incorporate the dialog domain knowledge [12]. In a human-machine dialog, uncertainty resulting from errors in speech recognition and ambiguities inherent in natural language may arise. Even though many approaches have been developed to address the problem of robust speech recognition in recent years, the accuracy of speech recognition systems degrades severely when the systems operate in an adverse environment. Inevitably, the SDS will encounter errors from ASR which thus result in miscommunication with the users [1, 2, 13].

Miscommunication handling is an important issue in the design of an SDS. In a dialog system, two types of miscommunication, non-understanding (Non-U) and misunderstanding (Mis-U) [14, 15], are generally encountered. Mis-U results from mismatched intentions between the speaker and listener, whereas Non-U occurs if the listener fails to obtain any interpretation or is not sufficiently confident to acquire an interpretation. In an SDS, Non-U results from the rejection of some keywords which will be filled in the semantic slots, and thus a response is required to request the user to rephrase the query in order to complete the semantic slots for intention understanding. On the other hand, for Mis-U, the sentence contains the recognized keywords which have been accepted and filled in the semantic slots but then are noticed as errors by the user in a later stage. Therefore, the user corrects the errors in the next dialog turn. Figure 1 shows an example for the occurrence of a Mis-U, in which the term "beef noodles" is misrecognized as "hot pot noodles." In the next dialog turn, the user tries to correct the content of the semantic slots so that the system can repair the misrecognized user's intention. Accordingly, the aim in designing a robust SDS is to make it error-aware of the recognized word sequence from ASR for Non-U recovery and Mis-U repair in order to achieve the real goal of a dialog.

Traditionally, the research community has focused on building an SDS for user intention detection and dialog management (DM). In spoken language understanding (SLU), dialog states (DSs) are the basic functional units [16] that describe the dialog behaviors in human-computer or human-human communication [17]. The features used to represent an utterance for DS detection include parts-of-speech (POSs) [18], semantic roles [19, 20], prosody [21, 22], and keywords [23]. With semantic analysis, statistical dialog management models, such as

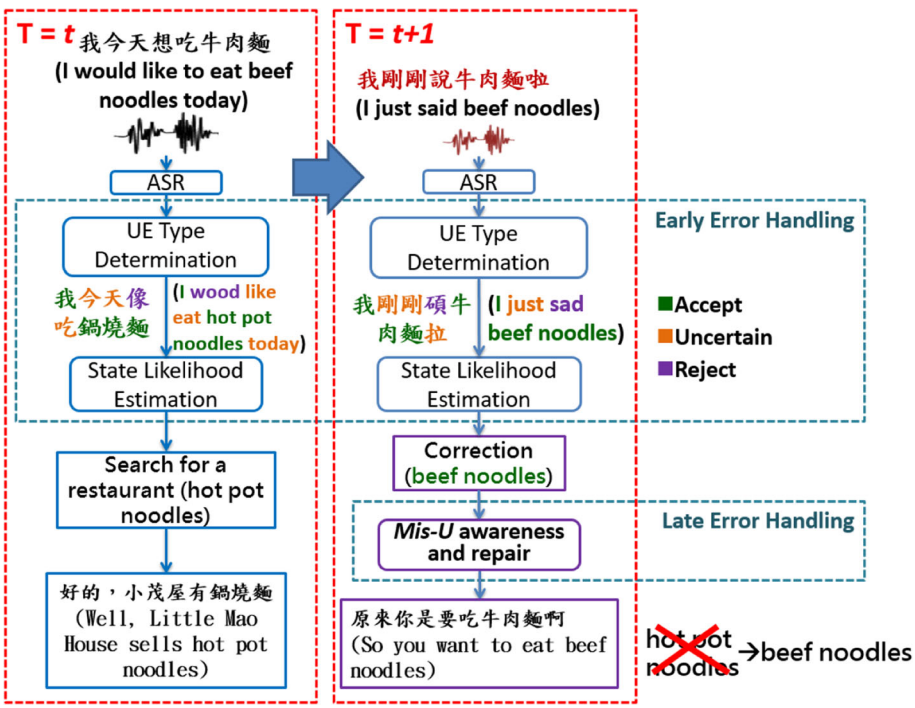

Fig. 1 An example for the occurrence of a *Mis-U* and the correction made by the user in the next dialog turn

weighted finite-state transducers (WFSTs) [24], Markov decision processes (MDPs) [25, 26], and partially observable MDPs (POMDPs) [8, 13], were proposed for stable dialog flow control, especially in goal-oriented SDSs. However, noisy speech, spontaneous speech with disfluencies [27, 28], or pronunciation variations [29] crucially affect the performance of SLU and DM. Intention detection based on only one-phase dialog response is not sufficiently robust for error-prone SDSs. On the other hand, error recovery with only repeated requests would also annoy the user and cost more turns significantly.

Handling of miscommunication and uncertainty in a dialog is not merely an individual task [30]. During the handling process, the system and the user share their individual knowledge and beliefs to establish what has been identified. Representative systems including Higgins (KTH) [30] and Let's Go! Public system (CMU) [31] endeavored to overcome potential dialog errors. Generally, the handling process consists of early and late error-handling stages. In early stage, confidence score obtained by ASR [30] based on intra-sentential [32] and inter-sentential information [33] is used to verify each recognized word and generate candidate sentences for further processing, such as asking the user to rephrase the query for Non-U recovery. Late stage performs repair actions against Mis-U issue. In addition, recent research has involved evaluating the understanding evidence (UE) for outcome analysis of the recognized words. Clark and Schaefer [34] listed five main types of positive UEs, such as continued attention and initiation of the relevant next contribution as shown in Table 1. These UE types are graded roughly from the weakest to the strongest. In Skantze's work [30], predefined UE types based on the ASR confidence score were employed to verify each word of the recognized word string.

This paper proposes a framework to model the dialog interaction process between a developed SDS and the users, as shown in Fig. 2. In the training phase, a simulated dialog environment for the restaurant reservation task was constructed for dialog corpus collection. The UE of each recognized word in the sentence from ASR output was firstly estimated. The UE, derived from the recognition confidence, is adopted for Non-U detection. As the input utterance is detected as Non-U, the system will request the user to rephrase the query for Non-U recovery. On the other hand, if the sentence contains the recognized keywords which have been accepted and filled in the semantic slots in previous dialog turn but then are noticed as errors by the user in the current turn, a Mis-U issue occurs. A set of Mis-U DSs is extracted using the partial sentences obtained by removing potentially misrecognized words from the input utterances. The partial sentences are organized, based on regular expressions, as a tree structure to tolerate the deletion or rejection of keywords resulting from misrecognition for Mis-U DS modeling. Based on the collected partial sentences, 28 Mis-U DSs are derived in this study. For the query with Mis-U, the system will correct the intention of the user's query based on the repair response from the user.

Table 1 The five types of understanding evidence (UE) [30, 34]

Type	Description
Continued attention	The hearer shows continued attention and remains satisfied with speaker's presentation.
Initiation of the relevant next contribution	The hearer starts in on the next contribution that would be relevant at a level as high as the current one.
Acknowledgement	The hearer says "uh huh," "I see," or nods.
Demonstration	The hearer demonstrates all or part of what he has understood.
Display	The hearer displays verbatim all or part of speaker's presentation.

Besides miscommunication cases, 37 Base DSs are defined in this study. For DS modeling, including Mis-U DSs and Base DSs, the linguistic features combined with word-level *n*-grams and syntactic rules parsed by the Stanford parser were extracted. Based on these features, a linguistic feature-by-DS (LFDS) matrix was built to describe the relationship between features and DSs. This study employed POMDP-based dialog state tracker as the dialog controller which monitors the probability distributions over the states in a Markov process. In the test phase, the user provided a query to the SDS. The UE of each word from the ASR output was estimated and used to determine whether to replace the rejected/uncertain word with the term "Filler." Next, linguistic features were extracted and sent to the LFDS matrix for DS detection. Then this study employed the POMDP-based DS detector to generate the response sentence. Finally, the response sentence was synthesized to give the speech output to the user.

The innovations of this paper are summarized as follows. First, few research focused on the error-aware issue which is the major problem in speech applications in recent years. Error handling, consisting of Non-U recovery and Mis-U repair, in spoken dialog applications is crucial for successful interaction. Second, we consider all error-tolerant sentence patterns based on partial sentence expansion to cover the recognition uncertainty resulting from an error-prone ASR. Finally, we organize these sentence patterns as an LFDS matrix to consider the relationship between the user utterance and the DSs. A POMDP-based dialog manager is then employed for dialog management.

The remainder of this paper is organized as follows. Section II describes the corpus collection and annotation. Section III introduces miscommunication handling for

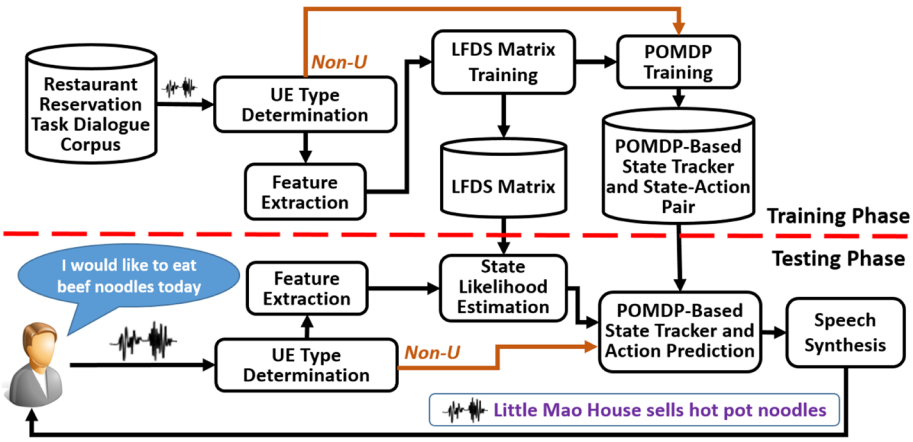

Fig. 2 The framework of the proposed spoken dialogue system

Non-U and Mis-U. Section IV describes the dialog state detection framework. Section V shows the experimental results and discussion. Finally, Section VI draws some discussion and conclusions, and suggestions for the future work.

2 Task corpus collection and annotation

For data analysis and system evaluation, we constructed a simulated SDS for collecting a query corpus for the task of food and restaurant information around the National Cheng Kung University (NCKU) campus. Figure 3 shows a Wizard-of-Oz (WoZ) framework for collecting the dialog data. Based on the keyword-spotting framework, the quantitative analysis on word verification was performed to estimate the UE of each recognized word. A partial sentence generation mechanism, trained from the recognized word sequences, was proposed to generate possible combinations to model the insertion and deletion errors obtained from an imperfect speech recognizer.

The dialog corpus was collected in three different ways—WoZ, web search, and human-system interaction. The WoZ method [30] was conducted for the collection of spoken dialogs. In the WoZ method, a human "wizard" mimics the functions of a system, either entirely or in part without the need for building a fully functional product. In web search, the restaurant information around National Cheng Kung University (NCKU) campus was obtained from the internet, including blogs, Google Map, and bulletin board systems, such as telnet://ptt.cc. We collected the sentences and then recorded the corresponding

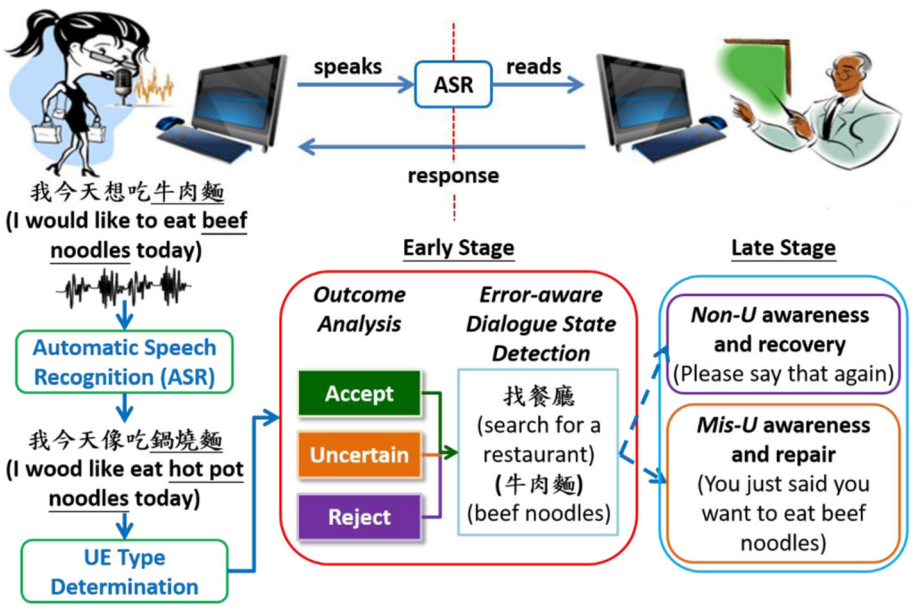

Fig. 3 The WoZ framework for miscommunication corpus collection

utterances. In human-system interaction, we traced the dialog processes, which were the essential materials for machine learning of a dialog manager. During the interaction, the unsuccessful dialog data were also observed, which resulted from imperfect ASR results.

Through manual labeling, the corpus was collected by 7 males and 4 females and finally we obtained 118 dialogs consisting of 1980 sentences totally. Stanford parser [35] was adopted to extract 682 syntactic rules. The collected queries with 432 keywords were obtained and classified into 26 semantic classes (SCs), including system service, restaurant address, food information, and greeting/ending, for semantic complexity reduction. For example, the sentences "北南山小館有賣牛肉麵嗎" ("Does Bei-Nan-Shan sell beef noodles?") and "三商巧福有賣排骨飯嗎" ("Does San-Shan-Chio-Fu sell roast-pork rice?") were transformed into "餐廳有賣菜色嗎" ("Does Restaurant_Name sell Food") and regarded as an argument of the DS "Query for Info (Name; Food)." Table 2 shows the predefined 37 Base DSs with the format comprising a predicate and optional argument derived based on the collected corpus. For the representation of a DS, the topic is within the parenthesis and the other words represent the predicate. Different occurrence frequencies refer to the variations in user behaviors. The three most frequently used DSs are "Appreciate," "Query for Restaurant Given (Food)," and "Goodbye". The DS "Query for Restaurant Given (Food)" shows that many people like to query the restaurant information; the others indicate that the users generally used "Appreciate," "Thank you," and "Goodbye" to end the dialog.

3 Miscommunication handling for Non-U and Mis-U

3.1 UE type determination

A Mandarin speech recognizer, using hidden Markov model toolkit (HTK) [36, 37], was implemented to construct the keyword recognizer and the possible outcomes were analyzed for miscommunication handling. A 39-dimensional feature vector consisting of 12-dimensional Mel-frequency cepstral coefficients (MFCCs) and one-dimensional log energy and their delta and acceleration features were adopted as acoustic features.

Based on the tied tri-phone sets extended from International Phonetic Alphabets (IPAs) [38], totally 602 models, each with 16 Gaussian mixtures, were constructed. SRILM [39] toolkit was used to obtain the related language model consisting of 234 uni-grams, 670 bi-grams, and 560 tri-grams. Combining with the language models, the constructed ASR system achieved an average word accuracy of 87.6% for the experiments in this study.

For Non-U and Mis-U detection, the proposed approach utilized the predefined UE types [30] to verify all the candidates in a recognized word sequence. The *z*-

Table 2 The predefined 37 Base DSs with predicate and optional arguments

No.	DS	No.	DS
1	System service consulting	20	Query for info (premium)
2	Query for restaurant (type)	21	Query for info (name; premium)
3	Query for restaurant (food)	22	Query for info (service)
4	Query for restaurant (type; evaluation)	23	Query for info (name; service)
5	Query for restaurant (food; evaluation)	24	Query for info (address)
6	Query for restaurant (address)	25	Query for info (name; address)
7	Query for restaurant (address; type)	26	Query for info (type)
8	Query for restaurant (address; food)	27	Query for info (name; type)
9	Query for restaurant (address; business hour)	28	Query for info (food)
10	Query for restaurant (address; service)	29	Query for info (name; food)
11	Query for restaurant (address; premium)	30	Query for info (price)
12	Query for restaurant (business hour)	31	Query for info (food; price)
13	Query for restaurant(service)	32	Query for info (name; food; price)
14	Query for restaurant (premium)	33	Appreciate
15	Query for info (name)	34	Good-bye
16	Query for info (phone)	35	Greeting
17	Query for info (name; phone)	36	Yes
18	Query for info (business hour)	37	No
19	Query for info (name; business hour)		

score [40] was further employed to normalize each candidate word's score, defined as follows.

$$z(w) = \frac{f(w) - \mu(w)}{\sigma(w)} \quad (1)$$

where $f(w)$ is the recognition score of the word w, $\mu(w)$ is the mean recognition score of all instances of the word w in the corpus, and $\sigma(w)$ is the standard deviation of all instances of the word w in the corpus. The quantity z represents the distance between the raw score and the population mean normalized by standard deviation. The quantity z is negative when the raw score is below the mean, positive when above.

The type of UE is determined by comparing the normalized recognition score of w using thresholds θ_1 and θ_2 defined in Eq. (2).

$$UE(w) = \begin{cases} \text{Accept} & , \text{ if } z(w) > \theta_2 \\ \text{Uncertain} & , \text{ if } \theta_1 < z(w) \leq \theta_2 \\ \text{Reject} & , \text{ if } z(w) < \theta_1 \end{cases} \quad (2)$$

where θ_1 is defined as two standard deviations below the mean $\mu(w)$ and θ_2 is the mean value $\mu(w)$. Figure 4 shows the mapping between the distributions of the verified words' scores and the UE types.

3.2 Non-U recovery

As Non-U results from the rejection of some keywords which will be filled in the semantic slots, when the average z-score of the keywords of an input utterance is below threshold θ_1, the SDS will identify the input utterance as Non-U and thus ask the user to rephrase the query in order to complete the semantic slots for intention understanding. For example, in this task, the SDS could respond to the user with the sentence "可以請您再說一次嗎" ("Please say that again") or "我聽不懂你說什麼" ("I do not understand what you say") to ask the user to rephrase the query for Non-U recovery.

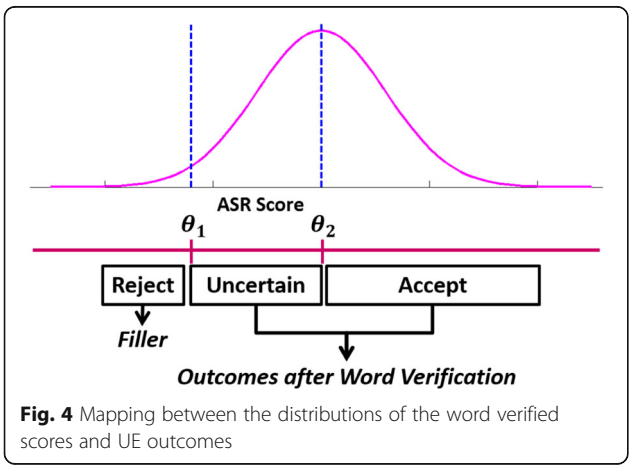

Fig. 4 Mapping between the distributions of the word verified scores and UE outcomes

3.3 Generation of partial sentences for Mis-U DS modeling

Even though misrecognized words may appear in any position of a given utterance, the partial fragments of the utterance passing the verification may have useful information for partial intention detection. Motivated by the idea, we attempt to extract the partial information from these parts to recover errors for DS modeling and detection. The idea is based on the words with UE types categorized as "Uncertain". Two types of sentences with uncertain recognition confidence or recognition errors are used to characterize the DSs using partial information of the recognized sentence.

(i) Sentences with potential recognition errors: The sentences with some content words detected as "Uncertain" are used to construct the partial sentence trees for characterizing the DSs with error-tolerant ability.
(ii) Sentences containing the words for repairing the semantic slots which may be incorrectly accepted in previous dialog turn: These sentences are used to model the DSs with repair ability.

In this study, we assume a sentence can be represented as a word sequence containing at least one functional word (*FW*) and other optional words (OWs). *FW*s represent the predefined semantic classes or keywords, and *OW*s are optional and could be omitted, e.g., "了(le)" and "吧 (ba)"

$$S = OW_1, OW_2, ..., OW_j, FW, OW_{j+2}, ..., OW_{NO+NF} \quad (3)$$

$$PS_i = OW^*(FW\ \ OW^*)^+ \quad (4)$$

where *NO* is the number of *OW*s and *NF* is the number of *FW*s. In Eq. (4), "*" is the Kleene star and "+" is the Kleene plus in a regular expression [41]. *FW* includes certain semantic classes to characterize a DS. These fragments are extracted from the recognized word sequence and used to simulate the imperfect recognition output. A tree structure, named partial sentence tree (PST) as shown in Fig. 5, is constructed by removing possible errors in the recognized sentence for a certain DS [42].

The basic idea of a PST is to replace unreliable word hypotheses with "Filler." In the PST, the word with "Reject" type was replaced with "Filler"; the word with "Uncertain" type could be either retained or replaced with "Filler"; and the word with "Accept" type will be retained in the PST without replacement. As a result, the constructed PST can cover the partial sentences containing all combinations of potential recognition errors. In a PST, a path from the root node to a leaf node represents a partial sentence. In Fig. 5, the misrecognized

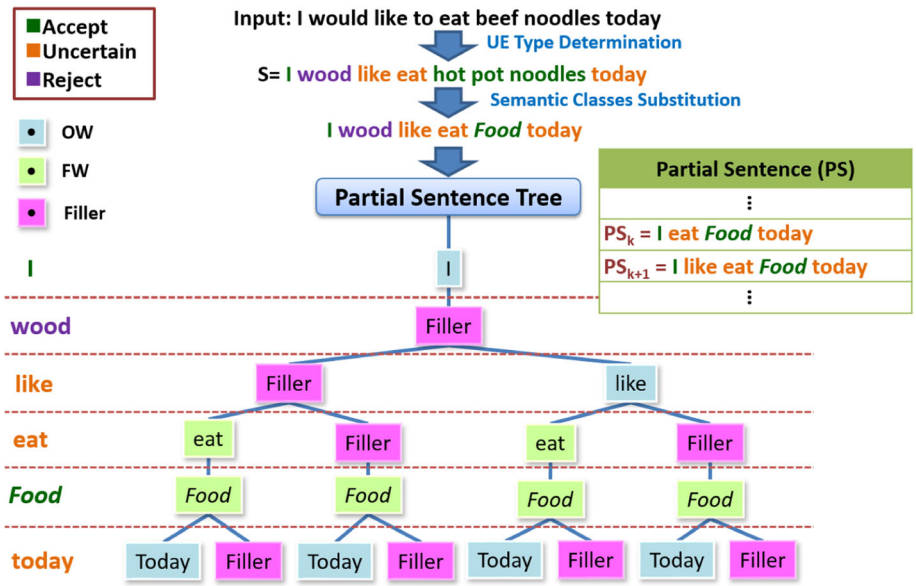

Fig. 5 An example for error-tolerant sentence generation based on partial sentence expansion

sentence "I wood like eat hot pot noodles today" was converted into the partial sentence pattern "I wood like eat Food today" after semantic class substitution. As the word "wood" is misrecognized and is replaced with "Filler," eight partial sentences are generated to represent the corresponding DS, such as "I eat Food today" or "I like eat Food today."

As all the partial sentences for each DS have been collected, 28 Mis-U dialog states are generalized. These partial sentences are used to train the DS detector for Mis-U DS detection.

4 Dialog state detection

As Bohus described in [43], a misunderstanding occurs if the response of the system mismatches the user intention. In this situation, the user may deliver an utterance with a correction, which is syntactically/semantically different from an ordinary response. These sentences may include short sentences or sub-phrases which imply "re-mention" or "negation," such as "我剛剛是說/問 (I just said/asked)" and "我不是說/問 (I did not say/ask)." In Clark's study [14], when misunderstanding happens, people prefer to deliver a direct phrase for correction. That is, the repair sentences frequently focus on the misunderstood semantic slots. Most phrases for the correction can only be mapped to the desired DS that contains only one argument. Based on these observations, historical information is crucial to improve the performance of a dialog system [40, 44].

Figure 6 shows the block diagram of the proposed SDS. An input utterance is decoded into a word sequence $\hat{W}$ from an ASR. In miscommunication handling, each recognized word in $\hat{W}$ is assigned a UE type based on the z-score. When the average z-score of the input utterance is smaller than a threshold value θ, the input utterance is identified as Non-U. Otherwise, the recognized word sequence is used for dialog state detection. In this study, Mis-U DSs can be further divided into two categories: *Error-tolerant* DS and *Repair* DS. *Error-tolerant* DSs are defined as the DSs in which some content words in the input utterance are verified as "Uncertain," and therefore, the recognized word sequence is represented by the PST to cover the potential errors. On the other hand, *Repair* DSs are used to represent the sentences which are trying to repair the semantic slots which had been incorrectly accepted in previous dialog turn. When the SDS identifies the input utterance as *Repair* DS, the SDS will extract the content words of the new input to repair the incorrect semantic slots.

Because the confidence measures are not entirely reliable and the user's language usage is often unexpected, PST generation is utilized to expand the word sequence $\hat{W}$ to several partial sentences in order to cover potential recognition errors. Due to different language usages for a query, the collected partial sentences for each predefined DS is further clustered into several variant DSs based on the k-means clustering algorithm using the linguistic features of the partial sentences in the PST, including word n-grams and syntactic rules obtained from the Stanford parser.

This paper proposes a dialog state detection approach to the detection of Base DSs and Mis-U DSs given the query utterance U and the dialog historical information DS_H.

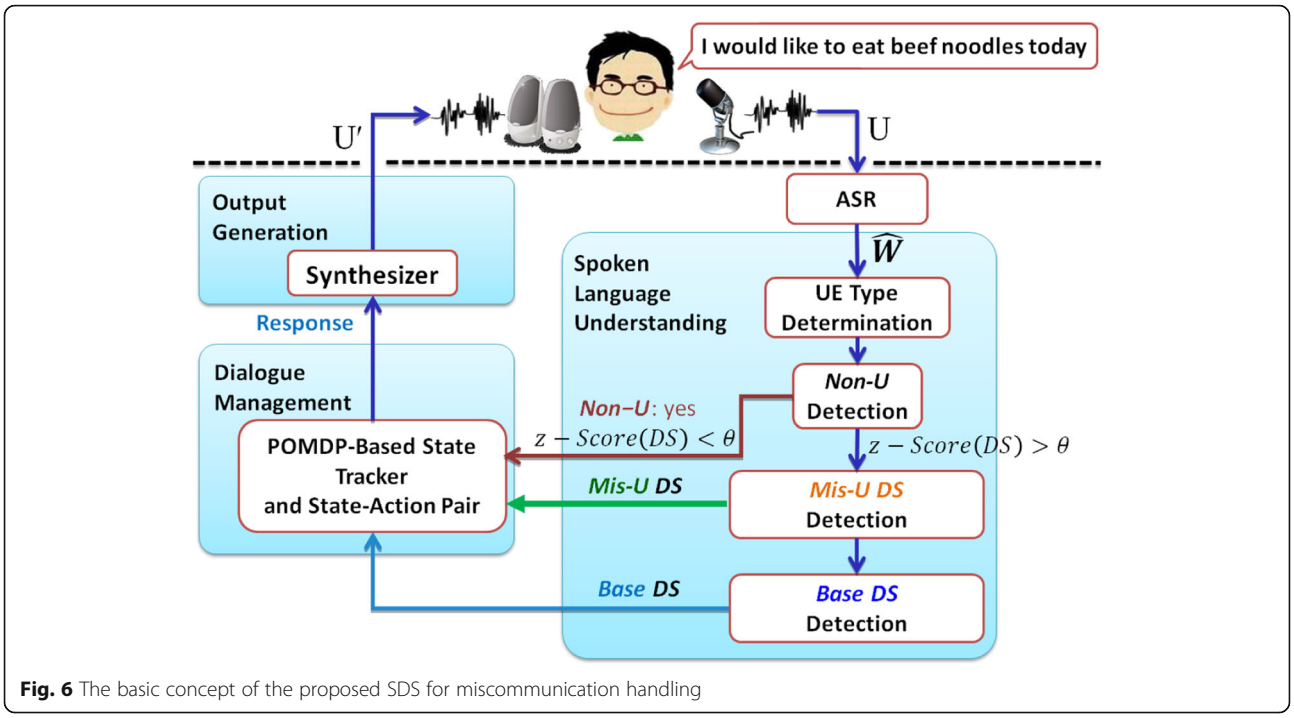

Fig. 6 The basic concept of the proposed SDS for miscommunication handling

$$DS^* = \arg\max_{DS_C \in \Omega_{DS}} P(DS_C|\mathrm{U}, DS_H) \tag{5}$$

where DS* represents the detected DS; Ω_{DS} represents the set of DSs, including Base DSs and Mis-U DSs; and DS_C is the C-th DS in the DS set. Given a word sequence $\hat{W} = \hat{w}_1,\ldots,\hat{w}_M$ decoded from the given utterance U, Eq. (5) is rewritten and derived based on Bayes' theory as follows.

$$\begin{aligned} DS^* &= \arg\max_{DS_C \in \Omega_{DS}} \sum_i P(DS_C, W_i|\mathrm{U}, DS_H) \\ &= \arg\max_{DS_C \in \Omega_{DS}, \widetilde{W} \in \Omega_W} P\left(DS_C, \widetilde{W}|\mathrm{U}, DS_H\right) \\ &= \arg\max_{DS_C \in \Omega_{DS}, \widetilde{W} \in \Omega_W} P\left(DS_C|\widetilde{W}, \mathrm{U}, DS_H\right) P\left(\widetilde{W}|\mathrm{U}, DS_H\right) \end{aligned} \tag{6}$$

Suppose that both Ŵ and U deliver the same information and the recognition result is independent of DS_H, the formula can be further rewritten as Eq. (7).

$$\begin{aligned} DS^* &\approx \arg\max_{DS_C \in \Omega_{DS},\ \widetilde{W} \in \Omega_W} P\left(DS_C|\widetilde{W}\right) P(DS_C|DS_H) P\left(\widetilde{W}|U\right) \\ &= \arg\max_{DS_C \in \Omega_{DS},\ \widetilde{W} \in \Omega_W} \frac{P\left(\widetilde{W}|DS_C\right) P(DS_C)}{P\left(\widetilde{W}\right)} P(DS_C|DS_H) P\left(\widetilde{W}|\mathrm{U}\right) \end{aligned} \tag{7}$$

Here, $P(\mathrm{DS}_C)$ and $P(\hat{W})$ have the same probability distribution for all DSs and thus can be ignored. Finally, the most likely DS* is determined according to the following equation:

$$DS^* \approx \arg\max_{DS_C \in \Omega_{DS}, \widetilde{W} \in \Omega_W} P\left(\widetilde{W}|DS_C\right) P(DS_C|\ DS_H) P\left(\widetilde{W}|\mathrm{U}\right) \tag{8}$$

where the probability $P(\hat{W}|\mathrm{DS}_C)$ is defined as the DS matching probability of the C-th DS with respect to the input word sequence $\hat{W}$; the probability $P(\mathrm{DS}_C|\mathrm{DS}_H)$ represents the conditional probability of the dialog historical information; and the recognition likelihood $P(\hat{\mathrm{W}}|\mathrm{U})$ is obtained from the ASR.

4.1 Feature extraction and language information matrix construction

In feature extraction, the n-grams and syntactic rules extracted by the Stanford parser are extracted as the linguistic features. These features comprise a 1464-dimensional n-gram feature vector and a 681-dimensional syntactic rule feature vector for each utterance. An example for the features of the query "Does this restaurant sell noodles? (這家餐廳有賣麵嗎?)" is given in Fig. 7. Besides, Stanford parser based on probabilistic context free grammar (PCFG) is utilized to extract the syntactic rules.

Figure 8 shows the constructed matrix [45] for error-tolerant DS detection. The co-occurrence relation between the linguistic features and each of the DSs are

The features of the query "**Does this restaurant sell noodles?** (這家餐廳有賣麵嗎?)"

		Features			Features
W^{Uni}	f_1	does	W^{Tri}	f_{16}	does-this-restaurant
	f_2	this		f_{17}	does-this-sell
	f_3	restaurant		f_{18}	does-this-noodles
	f_4	sell		f_{19}	this-restaurant-sell
	f_5	noodles		f_{20}	this-restaurant-noodles
W^{Bi}	f_6	does-this		f_{21}	restaurant-sell-noodles
	f_7	does-restaurant	Syntactic Rule	f_{22}	Root-SINV
	f_8	does-sell		f_{23}	SINV-VP
	f_9	does-noodles		f_{24}	VP-VBZ-does
	f_{10}	this-restaurant		f_{25}	SINV-NP
	f_{11}	this-sell		f_{26}	NP-DT-this
	f_{12}	this-noodles		f_{27}	NP-NN-restaurant
	f_{13}	restaurant-sell		f_{28}	NP-NN-sell
	f_{14}	restaurant-noodles		f_{29}	NP-NNS-noodles
	f_{15}	sell-noodles			

Fig. 7 A feature example for the query "Does this restaurant sell noodles? (這家餐廳有賣麵嗎?)"

estimated to construct a linguistic feature-by-DS (LFDS) matrix **Φ** as shown in Eq. (9).

$$\mathbf{\Phi} = \begin{array}{c} \\ f_1 \\ f_2 \\ \vdots \\ f_M \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{DS}_1 & \mathrm{DS}_2 & \cdots & \mathrm{DS}_N \end{array} \\ \begin{bmatrix} \phi_{1,1} & \phi_{1,2} & \cdots & \phi_{1,N} \\ \phi_{2,1} & \phi_{2,2} & \cdots & \phi_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{M,1} & \phi_{M,2} & \cdots & \phi_{M,N} \end{bmatrix} \end{array} \tag{9}$$

where each row is one of the linguistic features; each column represents a DS; $\phi_{m,n}$ implies the importance of the *m-th* feature f_m with respect to the *n*-th DS and is defined as follows.

$$\phi_{m,n} = (1-\varepsilon_m)P(f_m|DS_n) \tag{10}$$

The term $P(f_m|DS_n)$ indicates the importance of f_m to DS_n and is calculated as follows.

$$P(f_m|DS_n) = \frac{C(f_m, DS_n)}{\sum_{k=1}^{\mathbf{M}} C(f_k, DS_n)} \tag{11}$$

where $C(f_m, \mathrm{DS}_n)$ is the number of co-occurrences of f_m and DS_n in the corpus, and $(1 - \varepsilon_m)$ is an entropy-based measure for authenticity of f_m from the corpus and is regarded as a weight to $C(f_m, \mathrm{DS}_n)$. ε_m is obtained by Eq. (12).

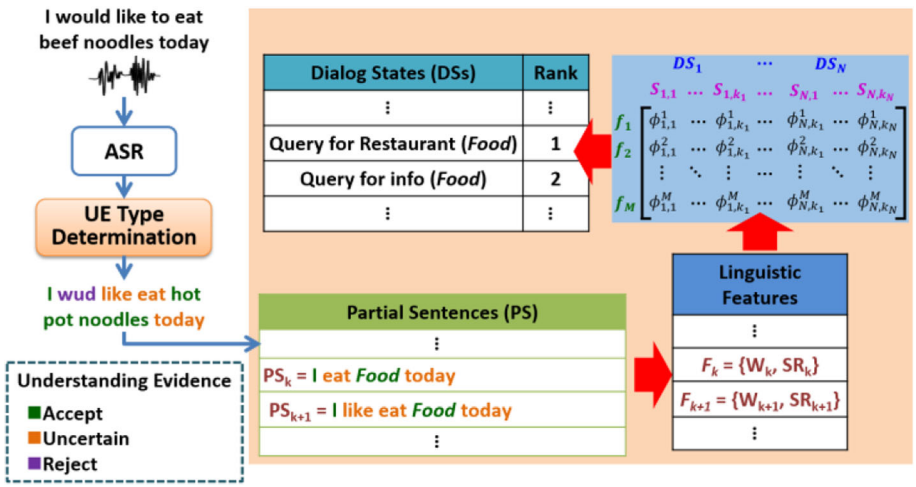

Fig. 8 An example of constructing the linguistic feature-by-DS (LFDS) matrix based on UE type

$$\varepsilon_m = -\frac{1}{\log \mathbf{N}} \sum_{n=1}^{\mathbf{N}} \frac{C(f_m, DS_n)}{\sum_{p=1}^{\mathbf{N}} C(f_m, DS_p)} \log \frac{C(f_m, DS_n)}{\sum_{p=1}^{\mathbf{N}} C(f_m, DS_p)} \tag{12}$$

4.2 Sentence clustering and linguistic feature transformation

During a dialog, users may deliver different utterances even when querying the same information. There would be various types of sentences belonging to the same dialog state. Different utterances which deliver the same query information can be used to enhance the flexibility for DS detection. The *k*-means clustering algorithm was employed to cluster the query sentences. For each DS, first, we manually select k unique sentences as the centroids of the sentence clusters. This algorithm partitions the sentences by maximizing the following function.

$$k^* = \arg\max_{k \in K} \sum\nolimits_{k=1}^{K} \sum\nolimits_{S_i} Similarity(S_i, S_k) \tag{13}$$

$$S_i \equiv \left[\delta_1^i,\ \delta_2^i, \ldots, \delta_m^i\right] \tag{14}$$

where S_i represents the binary feature vector of the i-th sentence; δ_m^i equals 1 if the m-th f_m belongs to the i-th sentence S_i; otherwise it is set to 0. The *cosine* similarity measure was used.

$$Similarity(S_i, S_k) = \frac{S_i \cdot S_k}{\|S_i,\| \cdot \|S_k\|} \tag{15}$$

The optimal number of clusters is empirically determined by using a cross-validation scheme. Through the clustering process, each state can be modeled by a set of sentence clusters $DS_n = \{S_{n,1}, S_{n,2}, \ldots, S_{n,k_n}\}$, where S_{n,k_n} represents the k_n-th sentence cluster in DS_n. The examples of the sentence patterns are shown in Table 3.

The latent semantic analysis (LSA)-based technique [45, 46] with entropy-based weighting scheme is employed to model the importance between features and DSs. In LSA, singular value decomposition (SVD) is performed to decompose the $\Phi_{m \times N}^{LFDS}$ into the product of the three matrices, $T_{m \times r}$, $S_{r \times r}$, and $T_{r \times N}$ with $r = \min(m, N)$. The truncated SVD models most of the significant underlying properties between linguistic features and DSs. This transformation yields a space with fewer dimensions of the row vectors, which is represented as a DS characteristic matrix (DCM), $\Phi_{m \times N}^{LFDS}$, where $\sigma < r$.

Table 3 Examples of sentence patterns

Sentences	Sentence patterns	DS
Do you provide its address information (請問你有提供它的地址資訊嗎)	f_1, f_2, f_3, f_5, f_7, f_9	Query for info (address)
Then please give me its address information (那請給我它的地址資訊)	f_5, f_7, f_{15}, f_{16}, f_{27}	Query for info (address)
where is it (它在哪裡)	f_4, f_6, f_{10}	Query for info (address)

4.3 Probability estimation

Given a sentence or word sequence $\hat{W}$, it can be represented as a feature vector.

$$F_{\mathbf{W}} \equiv (\delta_1,\ \delta_2, \ldots, \delta_m) \tag{16}$$

where δ_m is the quantity of the m-th linguistic feature obtained from $\hat{W}$. The DS matching probability $P(\hat{\mathbf{W}}|DS_t)$ is thus rewritten as

$$P(\hat{\mathbf{W}}|DS_t) \approx P(F_{\hat{\mathbf{W}}}|DS_t) \tag{17}$$

As each DS contains several sentence clusters, i.e., $DS_n = \{S_{n,1}, S_{n,2}, \ldots, S_{n,k_n}\}$, Eq. (17) is modified as

$$P(F_{\hat{\mathbf{W}}}|DS_n) = \max_{S_{n,K_n} \in \Omega_S} P(F_{\hat{\mathbf{W}}}|S_{n,K_n}) P(S_{n,K_n}|DS_n) \tag{18}$$

The equation can be expressed in Fig. 9. The term S_{n,k_n} stands for the k_n-th cluster in the n-th DS; $P(S_{n,k_n}|DS_n)$ is a weighting factor estimated using Eq. (19), where $C(S_{n,k_n})$ represents the number of sentences belonging to S_{n,k_n}.

$$P(S_{n,k_n}|DS_n) = \frac{C(S_{n,k_n})}{\sum_k C(S_{n,k_k})} \tag{19}$$

The *cosine* measure is employed to estimate the value of $P(F_{\hat{\mathbf{W}}}|S_{n,k_n})$.

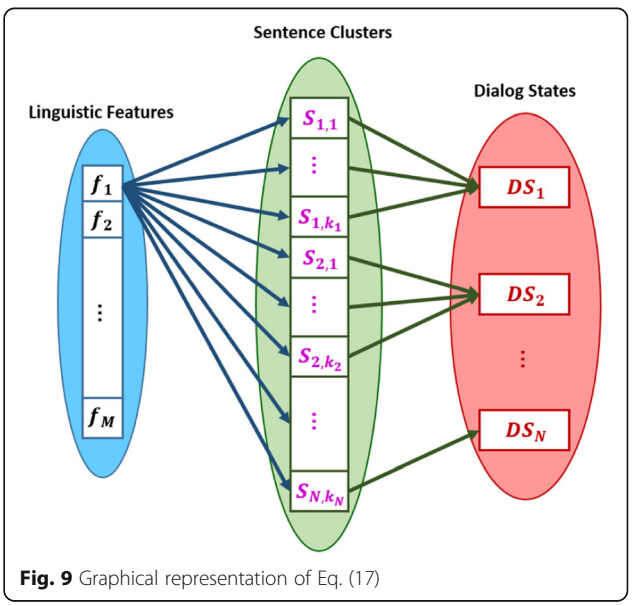

Fig. 9 Graphical representation of Eq. (17)

$$P\left(\mathbf{F}_{\hat{\mathbf{W}}}|S_{n,k_n}\right)=\frac{\mathbf{F}_{\hat{\mathbf{W}}}\cdot \mathrm{S}_{n,k_n}}{\left\|\mathbf{F}_{\hat{\mathbf{W}}}\right\|\cdot\left\|\mathrm{S}_{n,k_n}\right\|} \tag{20}$$

For estimating the dialog historical information, let DS_t denote the detected DS at the current turn (time t), and the historical state information be represented as $DS_H = \{DS_{t-1}, DS_{t-2},..., DS_1\}$. In this study, we assume that DS_t depends only on the previous correct DS confirmed by the speaker. A bi-gram model was employed to compute the probability $P(DS_t|DS_H)$. Combining with the word verification score as well as the probability $P(\hat{\mathbf{W}}|\mathbf{U})$, the most likely DS* can be obtained.

4.4 POMDP-based dialog manager

Formally, a POMDP [47] is defined as a tuple $\{S, A, T, R, O, Z, k, b_0\}$ where S is a set of states indicating the user's behavior; A is a set of system's actions; T is the transition probability $P(s_t|s_{t-1}, a_{t-1})$; R stands for the reward $r(s_{t-1}, a_{t-1})$; O is a set of observations the system can obtain from users; Z defines an observation probability $P(o_t|s_t, a_{t-1})$; k is a geometric discount factor where $0 \le \lambda \le 1$; b_0 is an initial belief state b_0 ($\boldsymbol{S}$).

The POMDP operates as follows. At each time step, a distribution over states is maintained by a belief function "b" with initial belief state b_0. We adopt $b(s)$ to indicate the probability of being in a specific state s. Based on b, the system takes an action $a \in A$, receives a reward $r(s, a)$, and transfers to a new unobserved state s', where s' depends only on s and a. Then the system receives an observation $o' \in O$ which is dependent on s' and a.

According to MDP assumptions, the latest state s_t is dependent on the last state s_{t-1} and system action a_{t-1}. However, in POMDP, s_t is not only dependent on s_{t-1} and a_{t-1} but also the past user action. The new observation allows the belief state b to be updated as follows:

$$\begin{aligned} b'(s') &= P\left(s'|o', a, b\right) \\ &= k\cdot P\left(o'|s', a\right)\sum_{s\in S}P(s'|s,a)b(s) \end{aligned} \tag{21}$$

Similar to MDP, a value function $V^*(b)$ is defined over states for optimal strategy selection according to the terms mentioned above and heuristic reward values:

$$V^*(b) = \max_{a\in A}\left[\sum_{s\in S}r(s,a)b(s)+\gamma\sum_{o',s,s'}p\left(o'|s',a\right)p\left(s'|s,a\right)b(s)V\left(b'\left(s'\right)\right)\right] \tag{22}$$

Referring to [48], Eq. (22) can be approximated by linear programming. The system will take action which leads to maximize $V(b)$ over the latest belief distribution.

In summary, the POMDP with error handling operates as follows. At each time step, a distribution over states is maintained by a belief function "b" with initial belief state b_0. We adopt $b(s)$ to indicate the probability of being in a specific state s. Based on b, if system observes an error, it executes an error-handling action $a \in A_{error}$, receives a reward $r(s, a)$, and transfers to a specific state $s' = Error\ state$ for next observation from the user; Otherwise, the system takes an action $a \in A$, receives a reward $r(s, a)$, and transfers to a new unobserved state s'. Then system receives an observation $o' \in O$ which is dependent on s' and a. Figure 10 illustrates the diagram of POMDP-based DS detection with error handling.

5 Experimental results and discussion

This work first addressed a simulated dialog corpus with text and read speech for the analysis of miscommunication phenomena. UE type determination was performed for Non-U and Mis-U detection followed by partial sentence generation. The Base DSs, Mis-U DSs, and Non-U were used to evaluate the proposed approaches. The fivefold cross-validation method was adopted; that is, 80% of the data were randomly selected and used for training, and the remaining 20% were used for testing.

5.1 Experiment on sentence clusters and LSA dimensionality

Generally, the sentence cluster number and the dimensionality in LSA are determined by the prediction risk. The most commonly used criterion which applies linear estimators for regression was employed. All the known analytical model selection criteria can be written as a function of the empirical risk penalized by the measure of the model complexity. In this paper, the generalized cross-validation method [49] was adopted as follows.

$$r(p) = (1-p)^{-2} \tag{23}$$

where $p = h_l/n$; r is a monotonically increasing function of the ratio p with degrees of freedom h_l and the training sample size n. After sentence clustering, a total of 494 clusters were obtained and used as an alternative approach to statistically characterize the sentence patterns. In determining the dimensionality of the reduced LSA space, the number of dimensions of the original feature vectors was 2145. This LSA transformation mapped the feature dimensions to the axes with large variations in the reduced LSA space. As shown in Eq. (23), the coverage rate

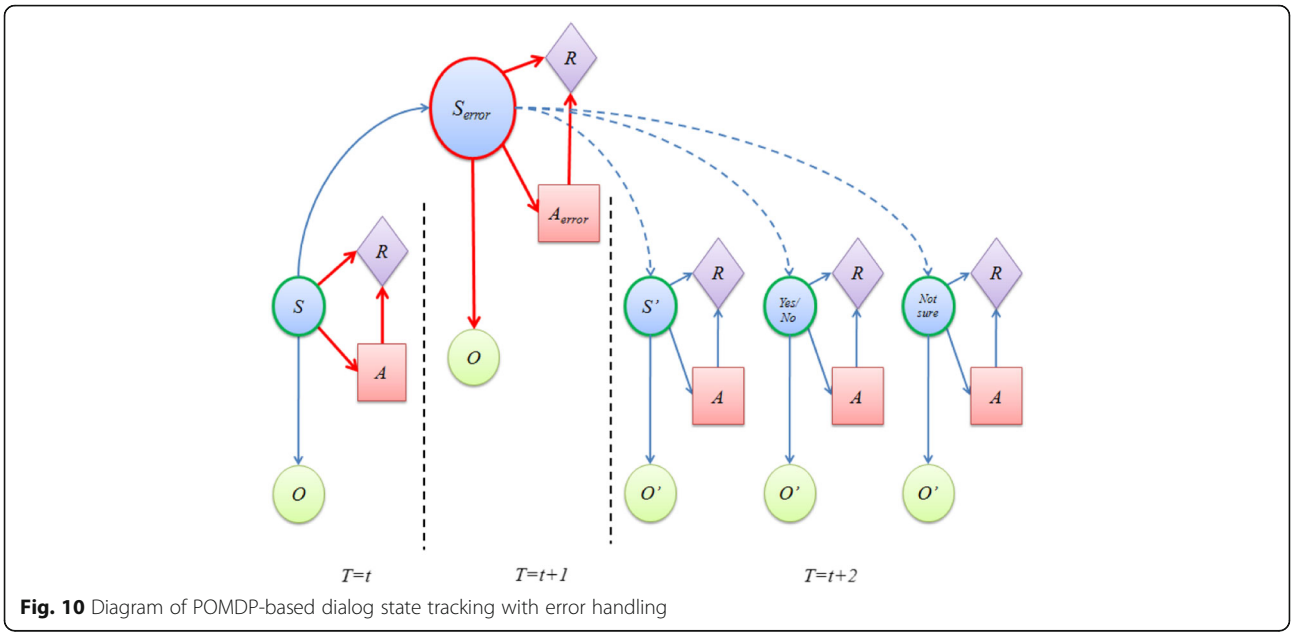

Fig. 10 Diagram of POMDP-based dialog state tracking with error handling

defined as the ratio of the cumulative variances of the selected features and the total features were considered as a statistical evaluation criterion. An alternative confidence level, ranging from 0.7 to 0.9, is often used. In this paper, the selection criteria involve the computational complexity and the coverage rate of the training corpus. A large number of features generally results in inconsistency in modeling the state distribution. In determining computational efficiency, a reduced dimension of 1716 at a coverage rate of 80% was selected to construct the DCM and was applied in the following experiments.

5.2 Experiments on LSA-based DS detection

Table 4 lists the evaluation results for the detection of Base DSs and Mis-U DSs. In Table 4, the baseline system for DS detection in this paper is based on keyword spotting. For comparisons, the DS detection performances for keyword spotting using the proposed approach with verification data, the proposed approach with/without sentence pattern clustering, and manual transcription are listed in Table 4. The experimental results show that the proposed approach achieved the detection rates of 83.95 and 81.46% for Base DS and Mis-U DS, respectively.

Table 4 Comparison of DS detection performances for the proposed approach using the verification data, the proposed approach with/without sentence clustering, and manual transcription

Average Detection Rate (%)					
	Keyword spotting	PA − SC + VD	PA − SC	PA + SC	Manual transcription
Base DS	50.16	65.30	75.39	83.95	94.57
Mis-U DS	51.96	72.59	80.68	81.46	92.69

PA − SC + VD proposed approach without sentence clustering using the verification data, *PA − SC* proposed approach without sentence clustering, *PA + SC* proposed approach with sentence clustering

Figure 11 illustrates the detection performance for each individual DS (DS $1 \sim 37$) using (1) keyword spotting, (2) the proposed approach using the verification data, and the proposed approach (3) without/(4) with sentence clustering. Clearly, joint estimation of LSA-based mapping between linguistic features and DSs as well as sentence clustering outperformed the individual approach and the traditional keyword-spotting method. Furthermore, the proposed approach can still show encouraging results, given the demand for sentence-based recognition of speech in variant dialog behaviors.

In this study, a linguistic feature-by-DS (LFDS) matrix was constructed to describe the relationship between the DSs and the user utterances. The linguistic features along with the word-level *n*-grams and syntactic rules were used to construct the LFDS matrix. As keyword spotting has been a technique for extracting targeted words in continuous speech sequence for years, and has also been verified as an efficient and effective method in SDSs [50], this study used the keyword spotting-based system [51] as the baseline to evaluate the proposed methods based on word-level models.

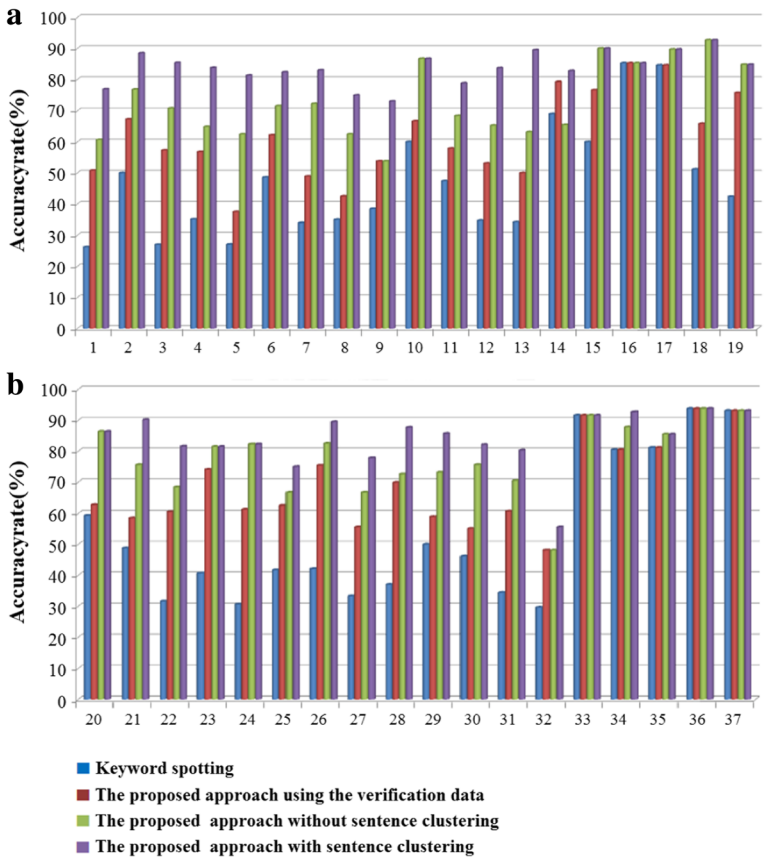

Fig. 11 Detection performance for each individual DS. **a** DS 1~19 and **b** DS 20~37 using (1) keyword spotting, (2) the proposed approach using the verification data, and the proposed approach (3) without/(4) with sentence clustering

Figure 12 shows the error analysis on the detection of Base DSs and Mis-U DSs. For Mis-U DS, in addition to the task domain keywords, such as "beef-noodles" or "Chang-Rong Road," the key sub-phrases, such as "我剛剛是說/問" ("I just said/asked"), "我不是說/問" ("I did not say/ask)," were also considered. These sub-phrases and their corresponding syntactic rules are beneficial to the detection of Mis-U DSs. Therefore, the number of mis-detections on Mis-U DSs was lower than that on Base DSs.

5.3 Practical evaluation of dialog management

Following Young's study [33], because of unreliable speech recognition, the conversation states can never be definitely known in a state-controlled process. The POMDP-based mechanism for dialog management with the proposed DS detection approach was constructed to evaluate the feasibility. The observation is defined as the score from DS detection.

$$P\left(o'|s',a\right)\cong P(DS_{o'}|DS_{s'},a) \tag{24}$$

where a is the system action, s' is the state, and o' is the observation which is dependent on s' and a. If a DS refers to Eq. (21) is rewritten as follows.

$$P(DS_{o'}|DS_{s'},a)\cong P(DS_{o'}|DS_{s'}) \tag{25}$$

However, if the DS refers to an error-handling action, the observation follows Eq. (24) because a reply from a user is a causal action based on system confirmation. Some rewards and their values, including (correct_by_sys, +20), (wrong_by_sys, −30), (terminate, +50), (welcome_out_of_starting run, −100), (task_complete, −15), (err_handle_correct_by_wizard, +20), (err_handle_wrong_by_wizard, −20), (expected_from_wizard, −100), and (unexpected_from_wizard, +100) were heuristically defined.

A set of the most commonly used response types (RTs), including "Accept," "Explicit Confirmation," "Ask Repeat," "Notify," "Re-prompt," "Ask_for_Different_Way," and "Replace" were adopted [52]. The response types (RTs) for Non-U recovery and Mis-U repair were generated by the following algorithm:

```
1  If ERROR = Non-U then
2    If UE (KW_i) = R then
       RT = Explicit Confirmation
3    Else
       RT = Ask Repeat or Notify or Re-prompt
4  Else If ERROR = Mis-U then
5    If Mis-U DS_t = Mis-U DS_t-1 or Mis-U DS_t = DS_t-1 then
       RT = ASK_in_diff_Way
     Else If pred (Mis-U DS_t) = pred(Mis-U DS_t-1) or pred(Mis-U DS_t) =
6    pred(DS_t-1) then
       RT = Give Alter
7  Else
     RT=Accept
```

To enhance the naturalness in the response of a dialog, more than one sentence template is provided for each response type. That is, the recovery or repair response was generated randomly by one response type template. From lines one to three, a Non-U recovery is provided for each condition. When the UE type of the i-th keyword KW_i is detected as "Reject," the "Explicit Confirmation" is provided. If a rejection occurs, one of the three RTs is provided, namely "Please repeat the utterance," "I do not know the keyword regarding the food," and "Please say the food again." From lines four to seven, a Mis-U repair aims to correct the undesirable values using new values or filling new slots.

The handling criterion follows three conditions. For Mis-U, if the detected Mis-U DS at the current turn is the same as Mis-U DS_{t-1} or Base DS_{t-1} in the previous turn, indicating that the predicates and arguments of Mis-U DS_t are all identical to that of the Mis-U DS_{t-1} and Base DS_{t-1}, this implies that the user provided the same sentence, but the system continually misunderstood. The system should reply with *ASK_in_diff_Way*, namely, "Please repeat the utterance in a different way." or "Could you please speak concisely?" If the two values KW_i^t and KW_i^{t+d} are received at the t-th and the $(t+d)$-th turns, respectively, and belong to the same i-th slot, substitution is applied based on their UEs; otherwise, *Give Alter* is employed for another confirmation. For

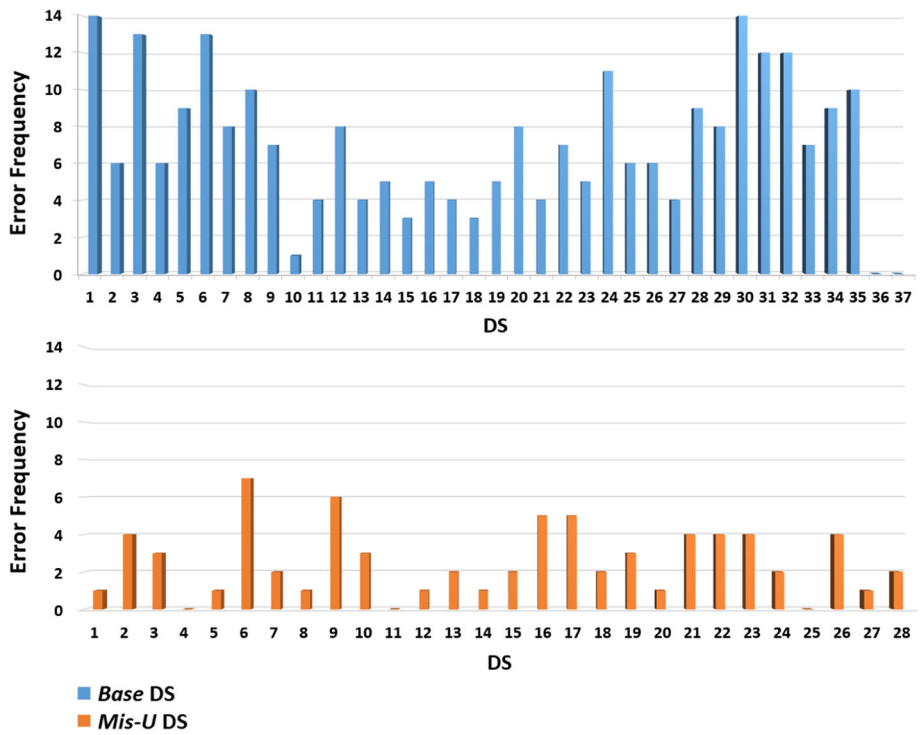

Fig. 12 Error frequency on the detection of Base and Mis-U DSs

line seven, DSs with different predicates indicate that the task should be completed. Therefore, the system regards this situation as *Accept*, which denotes another normal task execution.

Figure 13 shows the evaluation results and the comparison of the average task success rates for 20 simulated dialogs with three dialog strategies, ST1, ST2, and ST3, defined as follows.

(1) ST1: Slot filling using keywords. The system provides responses according to the recognized keywords from the user.
(2) ST2: POMDP + Baseline LSA + normal actions. Responses are based on POMDP by observing the Base DSs.
(3) ST3: POMDP + LSA with error awareness + handling strategy. Responses with error handling are based on POMDP by observing the DSs with error awareness from LSA.

Based on different recognition rate of ASRs, the number of turns required to complete the task will be different. Table 5 shows the comparisons of the average task success rates, average turn number, and average time spent for the 20 simulated dialogs with three dialog strategies.

The criterion of the average task success rate is defined as "$Task_j^i$ *is executed correctly*", where $Task_j^i$ represents the *j-th* task in the *i-th* dialog and is dependent on two situations:

(1) The task is successfully completed according to the DS in the same dialog session.
(2) The task is initially completed inaccurately, but then repaired with error handling, where these procedures are in the same sub-dialog.

The criterion of average task success rate is defined as follows:

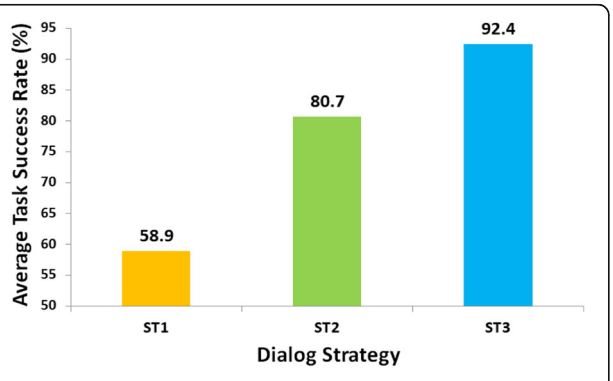

Fig. 13 Evaluation results and comparison of the average task success rates for 20 dialogs with three dialog strategies (ST1~ST3)

Table 5 Experimental results for different dialog strategies

Strategy	Average task success rate (%)	Average number of turns (average time spent)
ST1	58.9	17.68 (45.44 s)
ST2	80.7	12.47 (32.05 s)
ST3	92.4	13.73 (35.29 s)

$$R_{\text{Success}} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\text{Count}\left(\text{Task}_j^i \text{ is completed successfully}\right)}{\sum_{i=1}^{I}\sum_{j=1}^{J}\text{Count}\left(\text{Task}_j^i\right)} \tag{26}$$

For ST1, the dialog manager employed the ASR results directly. Therefore, those with misrecognition lead to incorrect slot fillings. For ST2, LSA alleviates the degraded ASR results and leads to robust DS detection. Furthermore, a dialog manager with more reasonable POMDP transitions achieved significantly superior results compared to that of ST1. Finally, ST3 with error handling obtained the best performance. In Table 5, the baseline system for dialog strategy decision is ST1 in which the dialog manager employed the ASR results directly. The experiments demonstrate that the proposed approaches not only improved dialog state detection rate but also handled potential dialog errors.

We have shown an error-tolerant framework for detecting DS from the developed corpus with 1980 utterances in 118 dialogs and 37 defined Base DSs. While it could be argued that the performance depends on the user queries and the simulated environment, further thought is that more training data and real observations are needed to improve the detection performance. In addition, as shown by the effects in the evaluation of mis-detection error analysis, the behavior of dialog turns as well as historical information plays important roles in this task. However, for simplifying the computation, the bi-grams was considered in the work. For investigating the performance of an SDS, a dialog management strategy and a robust model for inter-sentential analysis are also needed.

6 Conclusions

This work examined robust DS detection based on error awareness for miscommunication handling. The proposed approach aims to model the dialog behavior and detect the miscommunication situations by estimating the probabilities of DS matching and inter-sentential error correction. Mis-U awareness aims to avoid potential risk, such as out-of-condition dialog scenarios. The methods for sentence clustering and error-tolerant sentence generation could be applied to improve the

detection performance. LSA-based DS modeling with error awareness is also beneficial for DS detection. In the experiments, the average detection rates for slot filling using keywords were 58.9 and 80.7% for POMDP + Baseline LSA + normal actions. The proposed POMDP + LSA with error awareness + handling strategy achieved 92.4% task completion rate.

Experimental results reveal that the error-aware mechanism is robust in detecting a DS of interest in comparison with the traditional keyword-spotting scheme. These results also indicate that the proposed framework significantly improved the degraded performance of DS detection resulting from the error-prone speech recognizer. Besides, accented or noisy speech frequently appears in spoken language and results in poor ASR accuracy for the dialog system. From the evaluation point of view, an ASR with fair/poor recognition performance happens to be suitable to evaluate the ability of the proposed error-aware dialog act detection method. In the future, we will try to use a stronger ASR system (e.g., using the deep neural networks) to evaluate the usefulness of the proposed approach.

Acknowledgements
This research was supported in part by the Ministry of Science and Technology under Grant MOST105-2221-E-006-161-MY3.

Authors' contributions
C-HW, M-HS, and W-BL proposed the research methods, did the experiments, and wrote the main manuscript text, and all authors reviewed the manuscript. They all have equal contribution. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 14 December 2016 Accepted: 20 April 2017
Published online: 08 May 2017

References
1. JR Bellegarda, in *Spoken language understanding for natural interaction: the siri experience*, eds. by J. R. S. G. M. a. D. L. Mariani. Natural Interaction with Robots, Knowbots and Smartphones, (Springer, New York, 2014), pp. 3–14
2. KP Engelbrecht, in *A user model for dialog system evaluation based on activation of subgoals*, eds. J. Mariani, S. Rosset, M. Garnier-Rizet and L. Devillers. Natural Interaction with Robots, Knowbots and Smartphones (Springer, New York, 2014), pp. 363–374
3. J. a. J. F Henderson, in *Data-driven methods for spoken language understanding*, eds. O. Lemon and O. Pietquin. Data-Driven Methods for Adaptive Spoken Dialogue Systems (Springer, New York, 2012), pp. 19–38
4. S Kim, LF D'Haro, RE Banchs, JD Williams, M Henderson, J Williams, in *Proceedings of the 7th International Workshop on Spoken Dialogue Systems (IWSDS)*. The fourth dialog state tracking challenge, 1–14 2016
5. S Larsson, DR Traum, Information state and dialogue management in the TRINDI dialogue move engine toolkit. Natural Language Engineering **6**(3), 323–340 (2000)
6. D Bohus, AI Rudnicky, *Proceedings of Eighth European Conference on Speech Communication and Technology*, in *RavenClaw: Dialog management using hierarchical task decomposition and an expectation agenda*, 2003
7. T Paek, E Horvitz, in *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*. Conversation as action under uncertainty, (Morgan Kaufmann Publishers Inc, 2000), pp. 455–464
8. JD Williams, S Young, Partially observable Markov decision processes for spoken dialog systems. Computer Speech & Language **21**(2), 393–422 (2007)
9. D Bohus, A Rudnick, *Proceedings of the AAAI Workshop on Statistical and Empirical Methods in Spoken Dialogue Systems*, in *A 'K hypotheses+ other' belief updating model*, 2006
10. K Yoshino, T Hiraoka, G Neubig, S Nakamura, *Proceedings of the Seventh International Workshop on Spoken Dialog Systems (IWSDS)*, in *Dialogue state tracking using long short term memory neural networks*, 2016, pp. 1–8
11. V Zue, S Seneff, J Glass, J Polifroni, C Pao, T Hazen, L Hetherington, JUPITER: a telephone-based conversational interface for weather information. IEEE Transactions on Speech and Audio Processing **8**(1), 85–96 (2000)
12. M Henderson, *Proceedings of the First International Workshop on Machine Learning in Spoken Language Processing*, in *Machine learning for dialog state tracking: a review*, 2015
13. V Rieser, O Lemon, S Keizer, Natural language generation as incremental planning under uncertainty: adaptive information presentation for statistical dialogue systems. IEEE Transactions on Audio, Speech and Language Processing **22**(5), 979–994 (2014)
14. HH Clark, *Using language*, (Cambridge University Press, Cambridge, 1996)
15. M Henderson, C Matheson, J Oberlander, *Proceedings of SemDial 2012 (SeineDial): The 16th Workshop on the Semantics and Pragmatics of Dialogue*, in *Recovering from non-understanding errors in a conversational dialogue system*, 2012, pp. 1–8
16. H Bunt, in *Proceedings of THINK Quarterly*. Context and dialog control, **3**(1) 19–31 (1994)
17. R Prasad, M Walker, *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*, in *Training a dialog act tagger for human-human and human-computer travel dialogs*, 2002, pp. 162–173
18. R Levy, C Manning, *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, in *Is it harder to parse Chinese, or the Chinese treebank? (Association for Computational Linguistics*, 2003), pp. 439–446
19. C-H Liu, C-H Wu, *Proceedings of Interspeech*, in *Semantic role labeling with discriminative feature selection for spoken language understanding*, 2009, pp. 1043–1046
20. B Coppola, A Moschitti, G Riccardi, *Proceedings of NAACL-HLT*, in *Shallow semantic parsing for spoken language understanding*, 2009, pp. 85–88
21. H Wright, *Proceedings of ICSLP, Automatic utterance type detection using supra segmental features*, 1998
22. A Stolcke, N Coccaro, R Bates, P Taylor, C Van Ess-Dykema, K Ries, E Shriberg, D Jurafsky, R Martin, M Meteer, Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics **26**(3), 339–373 (2000)
23. T Kawahara, C-H Lee, B-H Juang, Flexible speech understanding based on combined key-phrase detection and verification. IEEE Transactions on Speech and Audio Processing **6**(6), 558–568 (1998)
24. C Hori, K Ohtake, T Misu, H Kashioka, S Nakamura, *Proceedings of Interspeech*, in *Dialog management using weighted finite-state transducers*, 2008
25. AG Barto, *Reinforcement learning: an introduction* (MIT Press, London, 1998)
26. S Singh, D Litman, M Kearns, M Walker, Optimizing dialog management with reinforcement learning: experiments with the NJFun system. Journal of Artificial Intelligence Research **16**, 105–133 (2002)
27. J-F Yeh, C-H Wu, Edit disfluency detection and correction using a cleanup language model and an alignment model. IEEE Transactions on Speech and Audio Processing **14**(5), 1574–1583 (2006)
28. C-H Wu, W-B Liang, J-F Yeh, Interruption point detection of spontaneous speech using inter-syllable boundary based prosodic features. ACM Transaction on Asian Language Information Processing **10**(6), 6–1-6:21 (2010)
29. W-B Liang, C-H Wu, Y-K Kang, *Proceedings of ISCSLP*, in *Recognition of syllable-contracted words in spontaneous speech using word expansion and duration information*, 2008, pp. 225–228
30. G Skantze, *Error handling in spoken dialog systems—managing uncertainty, grounding z and miscommunication*. (Doctoral Thesis, 2007)
31. A Raux, D Bohus, B Langner, AW Black, M Eskenazi, *Proceedings of Interspeech*, in *Doing research on a deployed spoken dialog system: one year of let's go! experience*, 2006

32. S Quarteroni, M Dinarelli, G Riccardi, *Proceedings of IEEE Workshop on Automatic Speech Recognition & Understanding*, in *Ontology-based grounding of spoken language understanding*, 2009, pp. 438–443
33. S Young, M Gasic, S Keizer, F Mairesse, J Schatzmann, B Thomson, K Yu, The hidden information state model: a practical framework for POMDP-based spoken dialog management. Computer Speech & Language **24**(2), 150–174 (2010)
34. HH Clark, EF Schaefer, Contributing to discourse. Cognitive Science **13**(2), 259–294 (1989)
35. The Stanford Parser: A statistical parser, http://nlp.stanford.edu/software/lex-parser.shtml. Accessed 8 Dec 2016
36. SJ Young, D Kershaw, J Odell, D, Ollason, V Valt-chev, P Woodl, *The HTK Book, version 3.4*. (Cambridge University Press, Cambridge, 2009)
37. S Young, G Evermann, D Kershaw, G Moore, J Odell, D Ollason, V Valtchev, P Woodland, in Handbook of the HTK book, vol 3 (Cambridge University Engineering Department, Cambridge, 2002), p. 175.
38. International Phonetic Association (IPA), *Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet*. (Cambridge University Press, Cambridge, 1999)
39. A Stolcke, *Proceedings of International Conference on Spoken Language Processing*, in *SRILM—an extensible language modeling toolkit*, 2002, pp. 901–904
40. S Hara, N Kitaoka, K Takeda, *Proceedings of LREC2010*, in *Estimation method of user satisfaction using N-gram-based dialog history model for spoken dialog system*, 2010, pp. 78–83
41. W-B Liang, C-H Wu, C-P Chen, *Proceedings of 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, in *Semantic information and derivation rules for robust dialog act detection in a spoken dialog system*, 2011, pp. 603–608
42. C-P Chen, C-H Wu, W-B Liang, Robust dialogue act detection based on partial sentence tree, derivation rule, and spectral clustering algorithm. EURASIP Journal on Audio, Speech, and Music Processing **1**, 1–9 (2012)
43. D Bohus, Error awareness and recovery in conversational spoken language interfaces. (Doctoral dissertation, SRI International, Carnegie Mellon University, Pittsburgh, CS-07-124, 2007)
44. S Hara, N Kitaoka, K Takeda, *Proceedings of 11th Annual Conference of the International Speech Communication Association (Interspeech)*, in *Automatic detection of task-incompleted dialog for spoken dialog system based on dialog act N-gram*, 2010, pp. 3034–3037
45. C-H Wu, Y-H Chiu, C-J Shia, C-Y Lin, Automatic segmentation and identification of mixed-language speech using delta-BIC and LSA-based GMMs. IEEE Transactions on Audio, Speech and Language Processing **14**(1), 266–276 (2006)
46. JR Bellegarda, Exploiting latent semantic information in statistical language modeling. Proceedings of the IEEE **88**(8), 1279–1296 (2000)
47. S Young, M Gasic, B Thomson, JD Williams, POMDP-based statistical spoken dialog systems: a review. Proceedings of the IEEE **101**(5), 1160–1179 (2013)
48. MT Spaan, N Vlassis, Perseus: randomized point-based value iteration for POMDPs. Journal of Artificial Intelligence Research **24**, 195–220 (2005)
49. V Cherkassky, F Mulier, *Learning from data: concept, theory, and methods*. (John Wuley & Sons, Inc., Hoboken, 1998)
50. M Vacher, B Lecouteux, JS Romero, M Ajili, F Portet, S Rossato, in *Proceedings of the International Conference on Speech Technology and Human-Computer Dialogue. Speech and speaker recognition for home automation: preliminary results*, (IEEE, Bucharest, 2015), pp. 1–10
51. C-H Wu, Y-J Chen, Multi-keyword spotting of telephone speech using a fuzzy search algorithm and keyword-driven two-level CBSM. Speech Communication **33**(3), 197–212 (2001)
52. K Ohtake, T Misu, C Hori, H Kashioka S Nakamura, in *Proceedings of Second International Symposium on Universal Communication (ISUC'08)*. Dialogue act annotation for statistically managed spoken dialogue systems, (IEEE, Osaka, 2008), pp. 416–422

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Immediate publication on acceptance
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com