Chen *et al. EURASIP Journal on Audio, Speech, and Music Processing* 2012, **2012**:13
http://asmp.eurasipjournals.com/content/2012/1/13

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Robust dialogue act detection based on partial sentence tree, derivation rule, and spectral clustering algorithm

Chia-Ping Chen[1], Chung-Hsien Wu[2*] and Wei-Bin Liang[2]

Abstract

A novel approach for robust dialogue act detection in a spoken dialogue system is proposed. Shallow representation named partial sentence trees are employed to represent automatic speech recognition outputs. Parsing results of partial sentences can be decomposed into derivation rules, which turn out to be salient features for dialogue act detection. Data-driven dialogue acts are learned via an unsupervised learning algorithm called spectral clustering, in a vector space whose axes correspond to derivation rules. The proposed method is evaluated in a Mandarin spoken dialogue system for tourist-information services. Combined with information obtained from the automatic speech recognition module and from a Markov model on dialogue act sequence, the proposed method achieves a detection accuracy of 85.1%, which is significantly better than the baseline performance of 62.3% using a naïve Bayes classifier. Furthermore, the average number of turns per dialogue session also decreases significantly with the improved detection accuracy.

1 Introduction

Spoken dialogue systems (SDS) are computer systems with which a user interacts through natural speech [1]. Services based on SDS have been deployed in a wide range of domains, from simple goal-oriented applications, such as DARPA Airline Travel Information System project for flight information [2], AT&T “How May I Help You?” for call routing [3], and systems for trip planning [4-6], to complex conversational applications, such as chatbot A.L.I.C.E. [7] and a variety of conversational agents using avatars [8].

The designer of an SDS often faces the following critical issues. First, with noisy speech or spontaneous speech with disfluency [9,10], abundant errors made by automatic speech recognition (ASR) can lead to misunderstanding or even pre-mature termination of a dialogue session (i.e., task failure). Second, the spoken language understanding (SLU) unit is often very expensive to develop, due to the manual annotation of certain features for semantic content. Examples of semantic features are part-of-speech tags [11], semantic roles [12,13], prosodic features [14], and keywords [15]. Third, the dialogue manager (DM) requires a sound dialogue strategy for management based on the state of a dialogue. Such a strategy could be quite complex in order to deal with all sorts of uncertainty, such as errors in ASR.

A dialogue act (DA) describes the purposes or effects of an utterance in a dialogue [16,17]. In principle, an utterance can convey multiple DAs. It is a succinct representation of the current intention of the speaker. DAs are closely related to speech acts (SA) [18], but they are specialized to dialogue systems [19]. While SAs are generic, DAs often vary from SDS to SDS. Since we are building an SDS, the notion of DA is more appropriate than SA to our study.

In this article, we describe an SDS with robust DA detection. Knowledge sources exploited include ASR confidence, semantic representation of ASR output, and the history of DA. First, the detrimental effects caused by ASR errors are abated by using partial sentence trees. Second, an unsupervised learning approach can determine data-driven DAs automatically, reducing annotation costs. Third, when DA can be reliably detected, the complexity of DM strategy can be significantly reduced. The motivation for focusing on robust DA detection is that the issues

* Correspondence: chunghsienwu@gmail.com
[2]Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan
Full list of author information is available at the end of the article

with ASR error, SLU cost, and DM complexity can be greatly alleviated.

A wealth of methods for DA detection have been introduced in the literature. The simplest and strongest “memorylessness” property basically assumes that the current (DA) is independent of the past. Thus, DA detection is based on a set of features derived from the current utterance. In this case, classification-based methods have been studied, including support vector machines (SVM) [12,20], naïve Bayes classifiers (NBC) [21-23], and multi-layer perceptrons (MLP) [14,21]. When the memorylessness assumption is relaxed, the dependence between past and current DAs has been modeled by *n*-grams [24,25], hidden Markov models [26,27], and Bayesian networks [28]. Recently, methods based on weighted finite state transducers (WFST) [4,29-31] or partially observable Markov decision processes (POMDP) [32-34] have been studied for DM.

Our method for DA detection is completely different. First, DAs are data-driven by clustering via the spectral clustering algorithm, with each cluster identified as a DA. The clustering happens in a space defined by derivation rules (DR). Classification of DA for unseen utterances is based on a novel derivation rule-dialogue act (DRDA) matrix, which is created by counting the occurrences of each DR in each utterance cluster. As a result, a column in the DRDA matrix represents a DA in the vector space spanned by DRs. As an example, in our system, the utterance How can I go to Anping-Fort by car? is mapped to DA-33 (Car_Destination, as listed in Table 1), and takes an action which leads to the generation of system response “The suggested line is that... “.

Table 1 List of dialogue acts

Numbers	DA	Numbers	DA
1	Greeting	2	Ending
3	Query_Service	4	Query_Spot
5	Query_Opening	6	Query_Introduction
7	Query_Contact	8	Query_Telephone
9	Query_Address	10	Query_Ticket
11	Query_Route	12	Query_Opening_Spot
13	Query_Introduction_Spot	14	Query_Contact_Spot
15	Query_Telephone_Spot	16	Query_Address_Spot
17	Query_Ticket_Spot	18	Query_Route_Spot
19	Query_Station	20	Query_Bus
21	Bus_From	22	Bus_Destination
23	Bus_From_Destination	24	Query_THSR
25	THSR_From	26	THSR_Destination
27	THSR_From_Destination	28	Query_TRA
29	TRA_From	30	TRA_Destination
31	TRA_From_Destination	32	Car_From
33	Car_Destination	34	Car_From_Destination
35	Route_From	36	Route_Destination
37	Route_From_Destination	38	Particle

The rest of this article is organized as follows. The basic framework of SDS is introduced in Section 2. The proposed robust DA detection method is stated in Section 3. Details of the implementation are described in Section 4. Experiments and discussion on the results are presented in Section 5. Lastly, concluding remarks are given in Section 6.

2 Spoken dialogue system

A dialogue session between a user and a statistical SDS consists of a chain of interleaving user turns and system turns, as illustrated in Figure 1. ASR outputs a string of words (or *N*-best list) $\mathbf{W}$ based on utterance U. SLU parses $\mathbf{W}$ and output a semantic representation. DM updates the belief on dialogue states, and accordingly decides the system’s action based on a policy. Natural language generation (NLG) converts system’s action to a surface representation in the textual form, which is passed to the text-to-speech (TTS) module for speech waveform generation. The cycle repeats when the user responds with the next utterance.

The ASR module turns user’s utterance into word hypotheses. A telephone-based SDS inevitably needs to deal with noisy speech and spontaneous speech, rendering the job of ASR module difficult. Furthermore, errors made by ASR may propagate along the system, making the jobs of other modules difficult. As a result, ASR accuracy is critical to the performance of SDS.

The SLU module, as depicted in Figure 2, converts ASR output into semantic representation. In the proposed system, the ASR output is first converted to a partial sentence tree (PST) [35] in the PST Construction block. The basic idea of PST is to replace unreliable word hypotheses by fillers. As a result, PST is less vulnerable to recognition errors. From PST, partial sentences are formed and parsed.

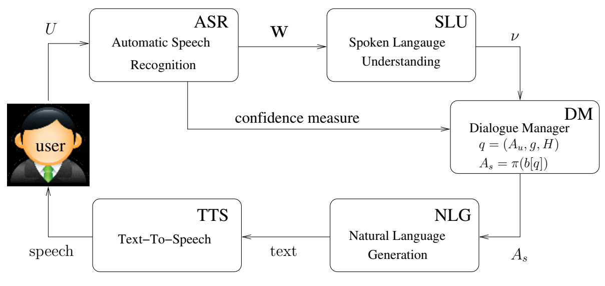

Figure 1 Block diagram of a spoken dialogue system. At turn t, the user utters U, which is recognized by ASR to be $\mathbf{W}$. ν is a semantic representation of user’s intended dialogue act. q is the dialogue state, where A_u is the hypothesized user’s dialogue act. g is user’s goal, and H is dialogue history. b is a distribution over dialogue states. A_s is the system’s action. The function π is called policy and it encodes the strategy of the dialogue manager.

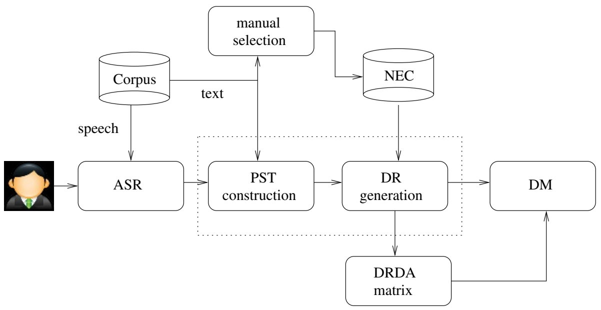

Figure 2 The spoken language understanding (SLU) module. During training, derivation rules are extracted based on partial sentence tree construction, and a derivation rule-dialogue act (DRDA) matrix is constructed. During testing, the trained DRDA matrix is used in DM for DA detection. Name entity class (NEC) inventory is referenced to convert certain words to word classes.

The parse results contain derivation rules (DR), which are extracted in the DR Generation block. The NEC (name entity class) inventory is referenced and certain words are replaced by word classes.

The core of an SDS is the dialogue manager. DM adopts sound strategy to keep dialogue sessions alive until they are successfully finished. An optimal action is taken at each turn based on the dialogue state, including user's goal, user's DA, and dialogue history. To cope with uncertainty, a belief on the states can be maintained, and the policy for taking action can be based on the belief.

3 Dialogue act detection

To infer dialogue act, a statistical model involving DA is required. The model assumption of the generation process for user's utterance is described as follows. Based on user's goal and the dialogue history, a user decides a DA, convert it into words, and produces an utterance. This is depicted in Figure 3. Note that each variable in the figure is indexed by turn t. However, to keep the notation and graph from being cluttered, we drop the subscript t. It is not difficult to see that the critical evidence to infer the current dialogue act should depend on ASR output, lexical items, and dialogue history. Thus, we can write

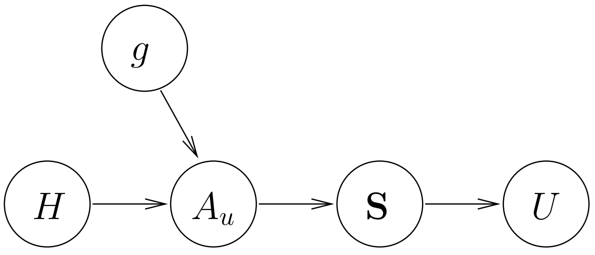

Figure 3 The generation process of user's utterance. In this graph, g is user's goal, H is dialogue history, A_u is user's intended dialogue act, $\mathbf{S}$ is the uttered sentence, and U is the acoustical observation. Note that the uttered sentence $\mathbf{S}$ and the recognition hypothesis of ASR $\mathbf{W}$ are different.

$$A_u^* = \arg\max_{A_u \in \Omega} \max_{\mathbf{W}} f(\mathbf{W},\ U) g(A_u,\ \mathbf{W})\ h(A_u,\ H), \qquad (1)$$

where $\Omega = \{A^1, . . ., A^q\}$ is the set of DAs. In (1), $f(\mathbf{W}, U)$ is called ASR score, $g(A_u, \mathbf{W})$ is called lexical scorem and $h(A_u, H)$ is called history score.

These scores are related to conditional probability functions. For the ASR score, we use the acoustic model and the language model in the ASR system. Specifically,

$$f(\mathbf{W},\ U) \approx p_{\mathrm{AM}}(U|\mathbf{W}) P^{\alpha}_{\mathrm{LM}}(\mathbf{W}), \qquad (2)$$

where $p_{\mathrm{AM}}(\cdot)$ is the acoustic model probability, $P_{\mathrm{LM}}(\cdot)$ is the language model probability, and α is the language model scale factor. For the history score, a back-off bi-gram model for DA sequence is used [4,30,31]. That is,

$$h(A_u,\ H) \approx Pr(A_t = A_u | A_{t-1}). \qquad (3)$$

Essentially, equation (3) models DA sequence as a Markov chain. We assume that the current user's DA depends on the history only through the previous user's DA. For the lexical score, a novel measure is proposed and the details are described in the following section.

4 Method for lexical score

One main contribution of this research is to demonstrate that a novel method for estimating lexical score $g(A_u, \mathbf{W})$ works quite well. The proposed method incorporates several steps, including partial sentence tree construction, derivation rule extraction, utterance representation in a vector space, the dialogue act set generation via spectral clustering, dialogue act representation using relative frequency weighted by normalized entropy, and finally a cosine distance measure between dialogue act and utterance. Taking the risk of being tedious, we describe the details of these steps in the following sections in order to make the overall procedure clear.

4.1 Construction of partial sentence tree

In an SDS, it is often beneficial to partition the vocabulary into a set of keywords $\mathcal{K}$ and a set of non-keywords $\mathcal{Q}$. Each word $w \in \mathcal{K}$ should be quite indicative of DA. Using $\mathcal{K}$ and $\mathcal{Q}$, the set of sentences with at least one keyword can be represented as

$$\mathcal{S} = \mathcal{Q}^*(\mathcal{K}\ \mathcal{Q}^*)^+ \qquad (4)$$

where $\mathcal{A}^*$ is the Kleene star (a.k.a. Kleene closure) of $\mathcal{A}$, and $\mathcal{A}^+$ is the Kleene plus of $\mathcal{A}$.

Given a sentence $s \in \mathcal{S}$, a partial sentence (PS) of s contains all keywords in s, while replacing some non-keywords in s by tokens called Filler. For a sentence with n non-keywords, there are 2^n PS's. These PS's can be compiled in a tree called partial sentence tree (PST). A path in PST from the root to a leave corresponds to a PS. The PST of sentence s is henceforth denoted by $\mathcal{T}_s$. For example, Figure 4 gives the PST for the sentence

$$s : \text{Where is the Anping - Fort} \tag{5}$$

In this example, Where and Anping-Fort are keywords, while is and the are non-keywords. The $2^2 = 4$ PS's embedded in the PST.

PST is a robust representation of ASR output. That is, even if some words are not recognized correctly, the semantics of an utterance can still be conveyed with the recognized keywords.

In the actual implementation, the ASR output is post-processed before PST construction. First, a word hypothesis, say w, is replaced by a Filler if the z-score [36] is below a threshold

$$z(w) = \frac{f(w) - \mu}{\sigma}, \tag{6}$$

where $f(w)$ is the recognition probability for word w, μ is the mean and σ^2 is the variance computed from all samples. In addition, recognized keywords are replaced by the named entity classes (NEC) or the greeting/ending classes, to have a compact representation.

4.2 Extraction of derivation rules

After PST construction, each PS in the PST is parsed by the Stanford parser (S-parser) [11]. Let the grammar of the S-parser be denoted as a 5-tuple [37]

$$G = (\mathcal{V},\ \Sigma,\ \mathcal{P},\ S,\ D), \tag{7}$$

where $\mathcal{V}$ is the set of variables, Σ is the set of terminals, $\mathcal{P}$ is the set of production rules, S is the sentence symbol, and D is a function defined on $\mathcal{P}$ for rule probability. In our implementation, a derivation rule (DR) is defined to be a derivation of the form

$$A \ \rightarrow \ B \ \rightarrow \ w, \tag{8}$$

where $A, B \in \mathcal{V}$ and $w \in \Sigma$. Note that equation (8) is a lexicalized rule. For illustration, parse results of the partial sentences are shown in Table 2. One can see that a lexical word in a PS produces a DR. Given a text corpus, a set of DRs $\mathcal{R} = \{R^1, R^2, \dots, R^l\}$ can be extracted and compacted.

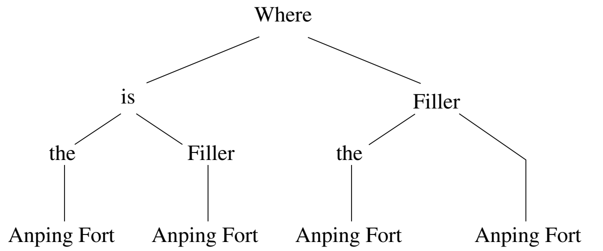

Figure 4 Construction of the partial sentence tree for the sentence where is the Anping-Fort. With 2 non-keywords, there are 4 partial sentences.

The motivation for using DR is to exploit the part-of-speech (POS) information. In particular, POS tags help to disambiguate noun-verb homonyms that occur quite often in Chinese.

4.3 Vector representation of sentences

Using each DR as a feature, we can represent a sentence s as a binary vector v_s, where

$$v_s(i) = \begin{cases} 1, \text{ if } R^i \in \mathcal{T}_s, \\ 0, \text{ otherwise}, \end{cases} \tag{9}$$

Table 2 Examples of the parse result (left) and the extracted derivation rules (right) corresponding to the four partial sentences in Figure 4

PS: where is the spot	
(Root	DR1: WHADVP WRB Where
(SINV	DR2: VP VBZ is
(FRAG	DR3: NP DT the
(WHADVP (WRB Where)))	DR4: NP NNP Spot
(VP (VBZ is))	
(NP (DT the) (NNP Spot)))	
PS: where is filler spot	
(ROOT	DR1: WHADVP WRB Where
(SBARQ	DR2: SQ VBZ is
(WHADVP (WRB Where))	DR3: NP NNP Filler
(SQ (VBZ is)	DR4: NP NNP Spot
(NP (NNP Filler) (NNP Spot)))))	
PS: where filler the spot	
(ROOT	DR1: WHADVP WRB Where
(FRAG	DR2: VP VB Filler
(WHADVP (WRB Where))	DR3: NP DT the
(VP (VB Filler)	DR4: NP NNP Spot
(NP (DT the) (NNP Spot)))))	
PS: where filler spot	
(ROOT	DR1: NP NNP Where
(NP (NNP Where) (NNP Filler) (NNP Spot)))	DR2: NP NNP Filler
	DR3: NP NNP Spot

where $\mathcal{T}_s$ is the PST for s. For example, the representative vector

$$v_s = [1\ 0\ 1\ 0]^T \tag{10}$$

means that R^1 and R^3 are used in $\mathcal{T}_s$, and that there are l = 4 derivation rules.

4.4 Generation of dialogue acts

We use a set of data-driven DAs to save the prohibitive cost of manual annotation. We apply the recently-proposed *spectral clustering algorithm* [38] to cluster utterances in the training set. The spectral clustering algorithm is chosen because a conventional clustering algorithm (e.g., k-means) is often sensitive to centroid selection (for initialization). After clustering, each cluster found is identified as a DA.

Our implementation of spectral clustering is outlined as follows. Suppose there are n utterances in the training set

$$\mathcal{D} = \{s_1,\ s_2,\ \ldots,\ s_n\}. \tag{11}$$

Each utterance is represented by a vector according to equation (9). From $\mathcal{D}$, we construct an $n \times n$ similarity matrix M, where the similarity $M_{kk'}$ between two utterances s_k and $s_{k'}$ is defined as the cosine measure between v_{s_k} and $v_{s_{k'}}$. The normalized Laplacian matrix of M is defined as

$$L \triangleq I - D^{-\frac{1}{2}} M D^{-\frac{1}{2}}, \tag{12}$$

where D is a diagonal matrix with entries

$$D_{kk'} = \delta_{kk'} \sum_{j=1}^{n} M_{kj}. \tag{13}$$

We find the eigenvectors of the q smallest eigenvalues of L. Note that the eigenvectors can be made orthonormal since L is real-symmetric. We put these eigenvectors in an $n \times q$ orthogonal matrix Q, and cluster the row vectors to q clusters. Each cluster is identified as a data-driven DA.

On a theoretical side, consider the conversion of M into a binary-valued matrix $\hat{M}$ via a threshold τ, i.e.,

$$\hat{M}_{kk'} = \begin{cases} 1, & M_{kk'} < \tau, \\ 0, & \text{otherwise}. \end{cases} \tag{14}$$

$\hat{M}$ can be regarded as the adjacency matrix of a graph $G = (\mathcal{N}, \mathcal{E})$, where node set $\mathcal{N}$ corresponds to $\mathcal{D}$, and edge set E corresponds to the non-zero entries in $\hat{M}$. It can be shown [38] that the multiplicity of the eigenvalue 0 for $\hat{L}$, the normalized Laplacian matrix of $\hat{M}$, equals the number of disjoint connected components in G, which can be identified as clusters in $\mathcal{D}$.

4.5 Derivation rule-dialogue act matrix

A cluster of utterances found via spectral clustering algorithm is identified as a DA. In our implementation, we use an entropy-based representation for DA. The representation of DA is described as follows. Let n_{ij} be the accumulated count that DR R^i occurs in the utterance cluster of A^j. From n_{ij}, a probability function of DA conditional on DR is defined as follows

$$\gamma_{ij} = \hat{P}(\mathrm{DA} = A^j | \mathrm{DR} = R^i) \triangleq \frac{n_{ij}}{\sum_{j'=1}^{q} n_{ij'}}, \quad i = 1, \ldots, l, \quad j = 1, \ldots, q. \tag{15}$$

The normalized entropy for the probability conditional on DR R^i is

$$\varepsilon_i = -\frac{1}{\log q} \sum_{j=1}^{q} \gamma_{ij} \log \gamma_{ij}, \quad i = 1, \ldots, l. \tag{16}$$

Note that $0 \leq \lfloor_i \leq 1$, and a DR R^i with a lower $\lfloor_i$ is more discriminative for DA. From equations (15) and (16), a matrix Γ of size $l \times q$ can be constructed with entries

$$\Gamma_{ij} = (1 - \varepsilon_i)\gamma_{ij}. \tag{17}$$

We call Γ the derivation rule-dialogue act (DRDA) matrix. The j^{th} column in Γ is a vector representation for a DA A^j in the vector space spanned by DRs.

4.6 Similarity between utterance and dialogue act

In our implementation, the lexical score $g(A_u, \mathbf{W})$ in equation (1) is decomposed into two terms

$$g(A_u,\ \mathbf{W}) \approx g_R(A_u,\ s) g_N(A_u,\ \mathbf{W}), \tag{18}$$

where $g_R(Au,\ s)$ is called DR score and $g_N\ (A_u,\ \mathbf{W})$ is called named entity score. For DR score, the following similarity measure is used

$$g_R(A_u = A^j,\ s) = \max_{\sigma \in \mathcal{T}_s} \frac{\mathbf{b}_\sigma^T \mathbf{a}_j}{|\mathbf{b}_\sigma|\,|\mathbf{a}_j|}, \tag{19}$$

where $\mathbf{b}_\sigma$ is the vector representation for PS σ in $\mathcal{T}_s$, and $\mathbf{a}_j$ is the vector representation for DA A^j (i.e., column j in DRDA matrix Γ). For named entity score, we use the naïve Bayes approximation

$$g_N(A_u = A^j,\ \mathbf{W}) = \prod_{\alpha \in \mathbf{W}} \nu(A^j,\ \alpha) \tag{20}$$

where α is a named entity. Note that $\nu(A^j,\ \alpha)$ is estimated from a training corpus by the relative frequency of α occurring in A^j.

5 Experiments and discussion

We evaluate the proposed method for dialogue act detection on an SDS for Tainan city tourist-information services.

5.1 Data collection

We adopt the setup utilized in [4,6] to collect the dialogue speech data. The data collection setup is shown in Figure 5, and an exemplar in the collected dialogue data is shown in Table 3. An operator play the role of SDS, which helps users to plan trips in Tainan. Twenty six male and eleven female subjects play the role of users. For our prototype system, users are asked to use utterances with single DA. Dialogue speech data is recorded in a lab environment, using 16,000-Hz sampling rate and 16-bit PCM format. There are 294 dialogues.

Two types of speech data are collected. The first type, called S-data, is from the operator playing the role of SDS. S-data contains travel information collected from on-line resources, such as Wikipedia and Google map. S-data set consists of 2, 653 utterances, with 317 different words. The second type, called U-data, is from subjects playing the role of users. U-data consists of 2, 636 utterances, with 297 different words. The vocabulary size is small as we have a domain-specific task. From U-data, 87 keywords corresponding to 28 named entity classes/ semantic classes and 796 derivation rules are obtained from the S-parser. Examples of the selected NECs and semantic classes are given in Table 4. The collected data contains sightseeing information, queries for the time schedules of two railway systems (Taiwan Railways Administration (TRA) and Taiwan High-Speech Rail (THSR)), and greeting/ending words in dialogues.

We use fivefold cross-validation method for system development. That is, the data is divided into five parts. In a round-robin fashion, four parts are used as training data, and one part is used as test data. We develop our system such that the average accuracy of DA detection over five test sets is optimal.

5.2 ASR module

The ASR module is an HTK-based Mandarin speech recognizer [39]. A syllable in Mandarin is modeled as the concatenation of an initial model and a final model. The acoustic model set includes 115 right-context-dependent initial models, 38 context-independent final models, 37 particle models, (e.g., EN, MA, OU), 47 syllable-level models for hyper-articulated speech, and 14 filler models (e.g., short pause, breathing, and footfall). Each initial model is a three state HMM, while each final model is a five state HMMs. The observation probability density of a state is a Gaussian mixture model (GMM) with no more than 32 components. The speech feature vector is composed of 39 components, including 12 mel-frequency cepstral coefficients (MFCCs), log energy, and the velocity and acceleration features. For real-world data with a variety of speakers, a reliable acoustic model is needed. Thus, an acoustic model set trained by the TCC-300 Mandarin corpus is adapted by the collected dialogue speech data via maximum-likelihood linear regression (MLLR). The lexicon contains 297 words. The bi-gram language model is estimated by SRILM toolkit [40].

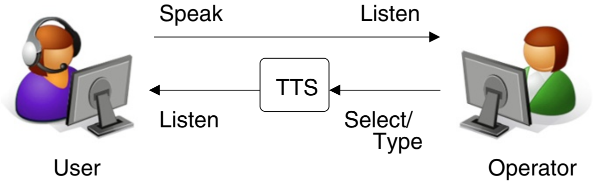

Figure 5 The environmental setting of data collection. The operator acts like an SDS, and the user acts like s/he is interacting with an SDS.

Table 3 The beginning part of a collected dialogue

Role	Dialogue act	Utterance
System	Greeting	Welcome
User	Query_Service	What service can you provide?
System	Ans_Service	I can provide the information of historic spot and timetable of railway
User	Query_Intro	Can you give an introduction to Anping-Fort?
System	Ans_Intro	Anping-Fort is also known as Fort Zeelandia It was first built by the Dutch in 1624 as ...

This is collected in the way illustrated in Figure 5

Table 5 shows the performance of ASR module with clean and simulated noisy speech. Note that the real-world scenario of noise corruption is applied in the collection of the noisy speech (footfall noise, human speech, or both). That is, a speaker stands in front of a microphone and the noise is played behind the speaker. From the results, we can see that the recognition accuracy does not severely degrade in the presence of noises behind a user.

5.3 The *z*-score threshold

Ideally, an effective threshold for *z*-score strikes a good balance between reliable recognition and keeping

Table 4 Examples of named entity classes (NEC) and semantic classes

NEC/semantic class	Named entities/words
City	Tainan, Taipei, Kaohsiung
Spot	Anping-Fort, Fort-Provintia, Sun-Moon Lake
Date	Today, Tomorrow, Yesterday
Time1	Morning, Noon, Afternoon
Time2	o'clock, hour, minute
Greeting	Welcome, Hello
Ending	Thanks, Bye

Table 5 Word accuracy rates of automatic speech recognition in clean and three noisy conditions

Conditions	Clean	Football	Human	Both noises
accuracy rate (%)	84.8	82.7	81.4	80.9

sufficient keywords for subsequent semantic representation in SLU. We analyze the rejected word hypotheses with z-scores below the threshold of 2 (which corresponds to the confidence level of 0.95), and find that only 3.4% of the keywords are incorrectly rejected. The threshold of 2 is therefore used. Such performance can be attributed to the fact that users often pause naturally before or after a keyword.

5.4 The number of dialogue acts

While generic speech acts are relatively well-defined, DAs are often specialized to particular domains and they need to be specified. In this research, since we adopt data-driven DAs by clustering, the number of DAs (clusters) become a critical parameter in system design. In order to decide this number, we investigate the system performance when it is varied. The detection accuracies are shown in Table 6. We can see that 38 DAs (q = 38) achieves the best performance[a]. Therefore, we use 38 DAs. To make more sense, each cluster is given an artificial but meaningful label (tag, name), as shown in Table 1. For example, Query_Introduction_Spot is assigned to the cluster formed by "queries of introduction to sightseeing spot".

5.5 Evaluation of feature sets

Just like the set of DRs, an alternative set of semantic features can be used as the coordinate axes to construct the corresponding vector set for $\mathcal{D}$. Applying spectral clustering, a matrix analogs to the DRDA matrix can be constructed according to the steps described in Section 4.

Including the proposed DR, 5 sets of features are investigated. In baseline, keywords are used as features. In NEC, named entity classes are used. In PS, partial sentences are used. In uwDR, derivation rules without normalized entropy weighting are used.

DA detection accuracies are summarized in Table 7. In this table, the column 40%-SIM means 40% of the words in the reference transcripts are retained, and similarly for 60%-SIM and 100%-REF. The recognition accuracy of ASR is 84.8% (15.2% word error rate), so we have a column of 84.8%-ASR. The middle columns are the results with simulated noisy speech, corrupted by footfall (football), human speech (human), and both noises (both).

Table 6 Accuracy rates of dialogue act detection with various numbers of DAs

Number of DAs	36	37	38	39	40
accuracy rate (%)	79.6	81.7	82.9	79.2	78.8

Table 7 Accuracy rates of dialogue act detection with various feature sets in various noisy conditions

	40%-SIM	60%-SIM	Football	Human	Both	84.8%-ASR	100%-REF
baseline	17.2	32.6	44.3	43.1	42.6	49.6	60.9
NEC	22.4	36.8	52.1	51.0	49.8	56.8	76.9
PS	29.8	49.2	75.2	74.6	73.5	76.2	91.1
uwDR	26.3	48.0	81.1	80.8	80.2	81.6	92.1
DR	26.3	47.4	82.3	81.9	81.6	82.9	93.3

In the case of 84.8%-ASR, we can see that NEC (56.8%) is better than baseline (49.6%), and that PS (76.2%) is better than NEC. The incorporation of uwDR (81.6%) and DR (82.9%) lead to further improvements. Thus, the difference between baseline and the proposed DR is very significant. We notice that an ambiguous Chinese word may correspond to different DAs with its different meanings. For instance, in open door and drive car, the words open and drive are the same word in Chinese. Using DRs helps disambiguation. For the cases of 40%-SIM and 60%-SIM, the results show clear improvement of NEC and PS over the baseline. Using DRs, however, does not further improve in these scenarios as the keywords are randomly discarded. We can see that recognizing the keywords is particularly important in highly adverse acoustic conditions. We also evaluate using the simulated noisy speech data in SDS. One can observe an interesting result that the performance of DR with the simulated noisy data and the clean data are very close. In PS, non-keywords are removed or replaced by Fillers. Thus, most of the partial sentences of simulated noisy speech are almost the same as those obtained from the clean speech.

5.6 Evaluation of the history score

The above results on DA detection are obtained without considering the dependency between DAs. Next, we evaluate the effectiveness of the history score. In order to balance the contribution of the lexical score and the history score, we generalize equation (1) to the following form,

$$A_u^* = \arg\max_{A_u \in \Omega} \; [g(A_u, \mathbf{W})]^{\beta_g} \; [h(A_u, \; H)]^{\beta_h}, \tag{21}$$

where $0 \leq \beta_g \leq 1$ is the weight of the lexical score, and $\beta_h = 1 - \beta_g$ is the weight of the history score.

A few comments on using equation (21) are in order. First, we note that when the ASR module outputs only one-best hypothesis, the maximization over $\mathbf{W}$ in equation (1) becomes trivial. It follows that the term $f(\mathbf{W}, U)$ can be dropped since it does not depend on A_u. In addition, as the values of $g(A_u, \mathbf{W})$ and $h(A_u, H)$ are in different ranges, simple linear combination may not work as one score can easily be dominated by the other. Therefore, we use the linear combination in the log domain, which is equivalent to the product in equation (21). In fact, a similar case based on the same consideration is the language model scale factor commonly used in ASR.

Table 8 shows the results of different β_h, and the best performance is achieved when β_h = 0.7. The evaluation results demonstrate that the dialogue history is informative.

5.7 Comparison with other methods

The performance of the proposed approach for DA detection is compared with other methods. In the NBC method, the keywords are used as the semantic features, and they are used in calculating DA probabilities. In the co-occurrence (co-oc) method, a priori algorithm [41] is used to calculate the co-occurrence of keywords in each DA. In the SVM and maximum entropy (ME) methods, a DA classifier is trained using keywords. In latent semantic analysis (LSA), the keyword-DA matrix is treated as a conventional word-document matrix, and then the LSA is applied. The results are listed in Table 9. We can see that the proposed approach achieves the best accuracy.

5.8 Evaluation on end-to-end measure

In addition to DA detection accuracy, we also conduct evaluation on end-to-end measures, i.e., from the start of a session to the end of the session. End-to-end measures are arguably better for performance evaluation as the ultimate goal of an SDS is to enable a user to complete a session correctly and quickly.

Three systems are evaluated, including the NBC, the proposed system (proposed), and the proposed system without using the history information (no history). Five subjects are recruited. Subjects perform exactly same task without knowing the order of the systems. This order is random for a test subject. A task is considered completed as soon as a subject acquires the appointed information. Table 10 shows the average dialogue turns per task of the evaluated systems. The proposed approach achieves the minimum of the average number of turns.

6 Conclusion

In this article, a robust dialogue act detection method using named entity classes, partial sentence trees, derivation rules, and entropy-based dialogue act-derivation rule matrix is investigated. Data-driven dialogue acts are created by the spectral clustering algorithm. Our implementation of a spoken dialogue system for tourist-information services incorporating the proposed method achieves 85.1% detection accuracy, outperforming a naïve Bayes classification based method (62.3%). It also reduces the number of dialogue turns per dialogue session on average. The results show that partial sentence tree and derivation rules are indeed succinct and informative features for dialogue act detection. Furthermore, spectral clustering is a successful method for automatic and unsupervised learning of dialogue acts from in-domain training data.

Table 8 Accuracy rates of dialogue act detection with various history score weights

Value of β_h	0.3	0.4	0.5	0.6	0.7	0.8	0.9
accuracy rate (%)	83.9	84.1	84.3	84.6	85.1	84.9	84.9

Table 9 Accuracy rates of dialogue act detection with five feature sets

Approaches	NBC	co-oc	SVM	ME	LSA	Proposed
accuracy rate (%)	62.3	62.6	75.8	76.3	78.6	85.1

Table 10 End-to-end measure of system performance evaluation.

	NBC	No history	Proposed
average number of turns	11	9.2	8.2

The average numbers of turns per dialogue session of three systems

Endnote

[a]Queries to 3 kinds of vehicles - bus, TRA, and THSR, are in different clusters when q = 38, but in the same cluster when q = 36. This partially explains the difference in performance between using 36 DAs and 38 DAs.

Author details
[1]Department of Computer Science and Engineering, National Sun Yat-sen University, 70 Lien-Hai Road, Kaohsiung, Taiwan [2]Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan

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

Received: 10 December 2011 Accepted: 3 March 2012
Published: 3 March 2012

References

1. N Fraser, in *Handbook of Standards and Resources for Spoken Language Systems*, vol. chap. 6, ed. by Gibbon D, Moore R, Winski R (Mouton de Gruyter, Berlin, 1997), pp. 564–564
2. PJ Price, Evaluation of spoken language systems: the ATIS domain, in *Proc the workshop on Speech and Natural Language*, (Hidden Valley, Pennsylvania, 1990), pp. 91–95
3. A Gorin, G Riccardi, JH Wright, How may i help you?. Speech Commun. **23**, 113–127 (1997). doi:10.1016/S0167-6393(97)00040-X
4. C Hori, K Ohtake, T Misu, H Kashioka, S Nakamura, Dialog management using weighted finite-state transducers, in *Proc INTERSPEECH-2008*, (Brisbane, Australia, 2008), pp. 211–214
5. J Liu, Y Xu, S Seneff, V Zue, CITYBROWSER II: a multimodal restaurant guide in Mandarin, in *Proc International Symposium on Chinese Spoken Language Processing*, (Kunming, China, 2008), pp. 1–4

6. T Misu, T Kawahara, Bayes risk-based dialogue management for document retrieval system with speech interface. Speech Commun. **52**, 61–71 (2010). doi:10.1016/j.specom.2009.08.007
7. R Wallace, The Artificial Linguistic Internet Computer Entity (A. L. I. C. E.) http://www.alicebot.org (2001)
8. J Cassell, T Bickmore, M Billinghurst, L Campbell, K Chang, H Vilhjálmsson, H Yan, Embodiment in conversational interfaces: rea, Proc the SIGCHI Conference on Human Factors in Computing Systems: the CHI is the limit, (Pittsburgh, Pennsylvania, 1999), pp. 520–527
9. JF Yeh, CH Wu, Edit Disfluency detection and correction using a cleanup language model and an alignment model. IEEE Trans Speech Audio Process. **14**(5), 1574–1583 (2006)
10. CH Wu, WB Liang, JF Yeh, Interruption point detection of spontaneous speech using inter-syllable boundary-based prosodic features. ACM Trans Asian Lang Inf Process (2010). 10, 6:16:21
11. R Levy, C Manning, Is it harder to parse Chinese, or the Chinese Treebank?, in *Proc 41st Annual Meeting on Association for Computational Linguistics (ACL)*, (Sapporo, Japan, 2003), pp. 439–446
12. CH Liu, CH Wu, Semantic role labeling with discriminative feature selection for spoken language understanding, in *Proc INTERSPEECH*, (Brighton, United Kingdom, 2009), pp. 1043–1046
13. B Coppola, A Moschitti, G Riccardi, Shallow semantic parsing for spoken language understanding, in *Proc Annual Conference of the North American Chapter of the Association for Computational Linguistics-Human Language Technologies*, (Boulder, Colorado, 2009), pp. 85–88
14. H Wright, Automatic utterance type detection using suprasegmental features, in *Proc International Conference on Spoken Language Processing*, vol. 4. (Sydney, Australia, 1998), pp. 1403–1406
15. T Kawahara, CH Lee, BH Juang, Flexible speech understanding based on combined key-phrase detection and verification. IEEE Trans Speech Audio Process. **6**(6), 558–568 (1998). doi:10.1109/89.725322
16. H Bunt, Context and dialogue control, in *THINK Quarterly*. **3**, 19–31 (1994)
17. R Prasad, M Walker, Training a dialogue act tagger for human-human and human-computer travel dialogues, in *Proc Annual Meeting of the Association for Computational Linguistics*, vol. 2. (Philadelphia, Pennsylvania, 2002), pp. 162–173
18. JL Austin, in *How to Do Things with Words*, ed. by Urmson JO, Sbis?á? M (Harvard University Press, Cambridge, MA, 1962)
19. A Stolcke, K Ries, N Coccaro, E Shriberg, R Bates, D Jurafsky, P Taylor, R Martin, Dialogue act modeling for automatic tagging and recognition of conversational speech. Comput Linguist. **26**(3), 339–373 (2000). doi:10.1162/089120100561737
20. G Tur, D Hakkani-Tür, L Heck, What is left to be understood in ATIS, in *Proc IEEE Workshop on Spoken Language Technologies*, (Berkeley, California, 2010), pp. 19–24
21. L Levin, C Langley, AL Donna Gates, D Wallace, K Peterson, Domain specific speech acts for spoken language translation, in *Proc the 4th SIGdial Workshop on Discourse and Dialogue*, (Sapparo, Japan, 2003)
22. S Grau, E Sanchis, MJ Castro, D Vilar, Dialogue act classification using a Bayesian approach, in *Proc Conference on Speech and Computer*, (St Petersberg, 2004), pp. 495–499
23. E Ivanovic, Dialogue Act Tagging for Instant Messaging Chat Sessions, in *Pro the ACL Student Research Workshop, Association for Computational Linguistics*, (Ann Arbor, Michigan, 2005), pp. 79–84
24. S Seneff, C Wang, TJ Hazen, Automatic induction of *N*-gram language models from a natural language grammar, in *Proc EUROSPEECH-2003*, (Geneva, Swiss, 2003), pp. 641–644
25. S Hara, N Kitaoka, K Takeda, Automatic detection of task-incompleted dialog for spoken dialog system based on dialog act N-gram, in *Proc INTERSPEECH-2010*, (Makuhari, Japan, 2010), pp. 3034–3037
26. K Ries, Hmm and Neural network based speech act detection, in *Proc IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. (Phoenix, Arizona, 1999), pp. 497–500
27. CH Wu, GL Yan, Speech act modeling and verification of spontaneous speech with disfluency in a spoken dialogue system. IEEE Trans Speech Audio Process. **13**(3), 330–344 (2005)
28. S Keizer, A Nijholt, Dialogue act recognition with Bayesian networks for Dutch dialogues, in *Proc the 3rd SIGdial workshop on Discourse and dialogue*, vol. 2. (Philadelphia, Pennsylvania, 2002), pp. 88–94
29. C Hori, K Ohtake, T Misu, H Kashioka, S Nakamura, Recent advances in WFST-based dialog system, in *Proc INTERSPEECH*, (Brighton, United Kingdom, 2009), pp. 268–271
30. C Hori, K Ohtake, T Misu, H Kashioka, S Nakamura, Statistical dialog management applied to WFST-based dialog systems, in *Proc IEEE International Conference on Acoustics Speech and Signal Processing*, (Taipei, Taiwan, 2009), pp. 4793–4796
31. K Ohtake, T Misu, C Hori, H Kashioka, S Nakamura, Dialogue acts annotation for NICT Kyoto tour dialogue corpus to construct statistical dialogue systems, in *Proc LREC2010*, (Valletta, Malta, 2010), pp. 2123–2130
32. JD Williams, S Young, Partially observable Markov decision processes for spoken dialog systems. Comput Speech Lang. **21**, 393–422 (2007). doi:10.1016/j.csl.2006.06.008
33. S Young, Still talking to machines (cognitively speaking), in *Proc INTERSPEECH2010*, (Makuhari, Japan, 2010), pp. 1–10
34. JD Williams, S Young, Scaling POMDPs for spoken dialog management. IEEE Trans Acoustic Speech Lang Process.**15**(7), 2116–2129
35. CH Wu, YJ Chen, Recovery from false rejection using statistical partial pattern trees for sentence verification. Speech Commun. **43**(1-2), 71–88 (2004). doi:10.1016/j.specom.2004.02.003
36. RJ Larsen, ML Marx, *An Introduction to Mathematical Statistics and Its Applications*, 3rd edn. (Prentice Hall, Lebanon, Indiana, USA, 2000). ISBN: 0139223037
37. D Jurafsky, JH Martin, *Speech and Language Processing*, 2nd edn. (Pearson Prentice Hall, New Jersey, 2009)
38. U von Luxburg, A tutorial on spectral clustering. Stat Comput. **17**(4), 395–416 (2007). doi:10.1007/s11222-007-9033-z
39. SJ Young, D Kershaw, J Odell, D Ollason, V Valtchev, P Woodland, *The HTK Book Version 3.4*, (Cambridge University Press, Cambridge, 2006)
40. A Stolcke, SRILM - an extensible language modeling toolkit, in *Proc International Conference on Spoken Language Processing*, (Denver, Colorado, 2002), pp. 901–904
41. R Agrawal, T Imielinski, AN Swami, Mining association rules between sets of items in large databases, in *ACM SIGMOD*, (Washington, D.C, 1993), pp. 207–216

doi:10.1186/1687-4722-2012-13
Cite this article as: Chen *et al.*: Robust dialogue act detection based on partial sentence tree, derivation rule, and spectral clustering algorithm. *EURASIP Journal on Audio, Speech, and Music Processing* 2012 **2012**:13.