AN EFFECTIVE APPLICATION OF CONTEXTUAL INFORMATION USING ADJACENCY PAIRS AND A DISCOURSE STACK FOR SPEECH-ACT CLASSIFICATION

KYUNGSUN KIM

Department of Computer Science, Sogang University
Seoul, 121-742, Korea
kksun619@gmail.com

YOUNGJOONG KO[1]

Department of Computer Science, Dong-A University
Pusan, 604-714,,Korea
yjko@dau.ac.kr

JUNGYUN SEO

Department of Computer Science & Interdisciplinary Program of Integrated Biotechnology, Sogang University
Seoul, 121-742, Korea
seojy@sogang.ac.kr

A speech-act is a linguistic action intended by a speaker. Speech-act classification is essential to the generation and understanding of utterances within any natural language dialogue system as the speech act of an utterance is closely tied to a user intention. Lexical information provides the most crucial clue for speech-act classification, and contextual information offers additional complementary clues. In this study, we concentrate on how to effectively utilize contextual information for speech-act classification. Our proposed model exploits adjacency pairs and a discourse stack to apply contextual information to speech-act classification. Experimental results show that the proposed model yields significant improvements in comparison with other speech-act classification models as well as a baseline model, which does not utilize contextual information.

Keywords: contextual information, adjacency pairs, discourse stack, shrinkage, speech-act classification, dialogue system

[1] Corresponding author

1. Introduction

Natural language dialogue systems are efficient tools that enable users to communicate with computers via a natural language dialogue interface. Because natural language interfaces are familiar and user-friendly, they have been used in several applications. A natural language dialogue system generally consists of a natural language understanding (NLU) module and a natural language generation (NLG) module. The NLU module converts an utterance into a representative form that the dialogue system can understand. The NLG module then converts the representation of an appropriate response, which is generated by the system based on an input utterance and built-in knowledge base within the dialogue system, into a natural language utterance.

To understand a natural language dialogue, the dialogue system must be able to determine the speaker's intention indicated through the speaker's utterances. Since a speech act is an intentional linguistic action, speech-act classification is essential for understanding an utterance within the context of a dialogue system. While researchers have developed many techniques for speech-act classification, they have found it difficult to infer a speech act from only a surface utterance. That is a reason why an utterance can represent more than one speech act if we do not consider contextual information [1, 2]. As shown in Table 1, the speech act of utterance (5) can be classified as "*Response*," "*Inform*," or "*Introducing-oneself*" within a surface analysis. To resolve this ambiguity, dialogue systems should analyze the context of an utterance. In this case, the choice of "*Response*" as the speech act of utterance (5) can be determined by considering the contextually previous utterance (4).

Table 1. An example dialogue annotated with speech acts.

Speaker	Korean	English	Speech act
(1) Agent	안녕하세요. 서울 호텔입니다.	Hello. This is Seoul Hotel.	Introducing-oneself
(2) User	가족이 4 명인데요.	I have four people in my family.	Inform
(3) User	방을 하나 예약 하려구요.	I want to reserve one room.	Request
(4) Agent	성함이 어떻게 되세요?	What is your name?	Ask-ref
(5) User	내 이름은 홍길동입니다.	My name is Kildong Hong.	**Response,** Inform, Introducing-oneself

In general, speech-act analysis has exploited multiple knowledge sources in the form of lexical, syntactic, prosodic and contextual information [3]. These sources have been typically modeled using various stochastic models. Conventional speech-act classification has relied on the words and syntax of utterances, whereas the ones of spoken dialogue systems, which require front-end speech recognition, have attempted to utilize prosodic information. Lexical information is used as a strong clue and prosodic information is as a complementary resource in the speech-act classification for the spoken dialogue systems.

Although contextual information can also be another significant clue for speech-act classification, it is true that the contextual information has not been studied systematically. Therefore, this paper focuses on how contextual information can be effectively applied to speech-act classification systems. A model is here provided to illustrate how the proposed system utilizes discourse structures and adjacency pair information as contextual information in lexical-based speech-act classification. The lexical-based speech-act classification is first executed using lexical features from a morphological analyzer, which consists of POS (part of speech) bigrams and content words; POS bigrams and content words represent the linguistic function and meaning of an utterance, respectively. The lexical-based classification has demonstrated better and robust performance than syntactic-based classification for speech-act analysis, because morphological analysis results have fewer errors than a syntactic parser [4]. As contextual information, adjacency pairs and discourse structures have the following properties in dialogue systems.

1. **Adjacency Pairs**: An utterance can be the first or second part of an exchange pair, such as request/accept, offer/accept, and question/answer pairs. Because both parts involve similar features, combining similar speech acts of a dialogue into a class and analyzing them as one class can help improve classification for all speech acts in the class. For this purpose, we first construct a two-level hierarchy for speech acts and employ the statistical technology of shrinkage to reflect the two-level hierarchy in our speech-act classification. In addition, since this method can generate considerable improvement in sparse speech acts, it would be a more appropriate method for use in surroundings with a data sparseness problem. For example, it is generally difficult to collect sufficient quantities of training dialogue examples, and, as a result, training data frequently includes a poorly balanced number of examples for certain speech acts.
2. **Discourse Structure**: While using previous utterances as contextual information can provide significant clues for classifying the speech act of a current utterance, most previous studies have regarded utterances located immediately prior to a current utterance as simply previous utterances, without consideration of their placement within a discourse structure such as a sub-dialogue. In order to detect and apply this sub-dialogue information effectively and efficiently, a discourse stack is developed for the proposed system. We can generate discourse rules using adjacency pair information, and a discourse stack is constructed and operated according to these

discourse rules. Using these discourse rules, a discourse stack enables us to effectively detect a previous utterance correctly by considering the sub-dialogue.

The proposed speech-act classification model achieved significant improvement in our experiments for speech-act classification, since it can effectively reflect contextual information into speech-act classification by utilizing the abovementioned adjacency pairs and discourse structures. In addition, a traditional spoken-dialogue system has a complicated architecture that consists of automatic speech recognition, natural language understanding, dialogue manager and natural language generation to generate a proper system response to user's utterance. In fact, a practical dialogue system requires a simple and robust architecture that can reduce human efforts to annotate training corpus and has fast processing time as a real-time systems. Even though the proposed model uses only low-level linguistic features including content words and POS tags, it obtained better performance than previous approaches with syntactic information by effectively using contextual information. It means that the model can be efficiently used in practical application areas because its training corpus can be easily constructed and scaled up and its architecture can be lighter than the conventional system.

The remainder of the paper is organized as follows. Section 2 presents other work related to this discussion. Section 3 provides a detailed explanation of the proposed speech-act analysis model. Section 4 discusses the experimental results, and the final section presents some concluding remarks.

2. Related Work

A dialogue is essentially a series of speaker turns. Utterances can be defined as the atomic subparts of a turn that accomplish one or more functions with respect to speaker interaction. Linguistics has identified several dimensions for the role of a sentence uttered in a dialogue: speech acts, turn management, adjacency pairs, overall organization and topics, politeness management, and rhetorical role [5]. These dimensions are not mutually exclusive, and speech acts among them play the most important role in detecting the function of an utterance. While there is not much agreement on the definition of a dialogue act, a dialogue act is generally considered a specialized speech act; one of the main inspirational sources for the tag sets of dialogue acts are speech acts, but dialogue acts are differently defined in different dialogue domains or systems [6]. After spoken dialogue systems became a commercial reality around 2000, the amount of research on dialogue acts has increased; dialogue acts have been gradually enriched with other possible functions in different domains for spoken dialogue systems, and a probabilistic integration of speech recognition using dialogue modeling has developed to improve both speech recognition and the accuracy of dialogue act classification [7-11].

Dielmann and Renals presented a framework for the automatic recognition of dialogue acts in multiparty conversations. This framework employed a generative probabilistic approach implemented through the integration of a heterogeneous set of technologies [7]. Rangarajan et al. presented a maximum entropy intonation model for dialogue act tagging that uses *n*-gram features of the normalized and quantized prosodic contour [8]. Laskowski and Shriberg defined a new set of features for dialogue act recognition in multiparty meetings, to aid in the detection of phenomena occurring at speaker turn edges [9]. They also proposed a framework for employing both speech/nonspeech-based (contextual) features and prosodic features and applied this framework to dialogue act segmentation and classification in multiparty meetings [10]. Stolcke et al. proposed a statistical approach for modeling dialogue acts in conversational speech. Their model detects and predicts dialogue acts based on lexical, collocational, and prosodic cues, as well as on the discourse coherence of a dialogue act sequence [11].

Although prosodic features can generally improve the performance of dialogue-act classification in spoken dialogue systems, lexical-based speech-act classification using contextual information provides a crucial evidence for dialogue-act classification. This type of classification is a fundamental and essential technique that utilizes domain-independent properties for the development of dialogue systems. Thus, we herein focus on how to apply contextual information to lexical-based speech-act classification.

Some previous studies on Korean speech-act classification have been based on rules extracted from a tagged dialogue corpus [1,12-14], while others have been based on statistical models learned from a tagged dialogue corpus [2,15-19]. The initial speech-act classification studies used rules extracted from a tagged dialogue corpus, such as linguistic rules and dialogue grammar.

Lee developed a two-step speech-act classification system using linguistic rules and dialogue flow diagrams; the first step in this model classifies surface speech acts, whereas the second step classifies deep speech acts [14]. Choi et al. proposed a statistical dialogue classification model that performs both speech-act classification and discourse structure analysis using maximum entropy [17]. This model automatically acquires discourse knowledge from a discourse-tagged corpus to resolve ambiguities. Lee and Seo classified speech acts by applying a bigram hidden Markov model (HMM) [18]. They used a forward algorithm to compute the speech act probabilities for each utterance. While computing speech-act probabilities to find the best path within the HMM, they encountered a sparse data problem, which they resolved by smoothing the probabilities using class probabilities with decision trees. Kim et al. proposed a neural network model and a method for extracting morphological features to classify Korean speech acts [15]. Their proposed neural network gave better experimental results than other models using comparatively high-level linguistic features. Errors occurring through the use of previously developed speech-act classification models result mainly from incomplete syntactic features and insufficient training data [18, 20]. To resolve the problems of previous work, the proposed method achieved the following two improvements. The proposed model first does not use syntactic and semantic features unlike previous work. We use lexical information and discourse features such as content words, POS tags and adjacency pairs, because morphological analyzers generally make much less errors than syntactic and semantic analyzers. In addition, we think that it is not easy to use syntactic and semantic analyzers in practical use because they need costly handcrafted knowledge and long analyzing time. Secondly, we use a shrinkage technique to get rid of the ill-balanced speech-act distribution problem. The shrinkage technique can compensate the ill-balanced distribution with speech-act hierarchy and increase overall speech-act classification performance.

3. The Proposed Speech-Act Analysis Model

The feature set of an utterance consists of lexical and discourse features. Contextual information can be applied to both of these feature sets. The application of contextual information in speech-act classification is based on adjacency pairs; the feature-weighting method for lexical features uses a speech-act hierarchy that is constructed based on adjacency pairs, and discourse features are generated using a discourse stack and discourse rules based on the adjacency pairs. Figure 1 presents an overview of the proposed speech-act classification model. The proposed model is composed of two modules: a lexical feature extraction and weighting module, and a discourse feature generation and weighting module. The former module first extracts lexical features based on POS tag information from a morphological analyzer and then estimates the feature weight using a speech-act hierarchy; lexical features consist of content words that reflect the meaning of an utterance and POS bigrams that reflect linguistic relationships between two consecutive POS tags. The latter module generates discourse features by analyzing relationships between the current and previous utterances: the speech act of a contextually previous utterance and discourse structural information. These discourse features are weighted by a simple binary weight scheme.

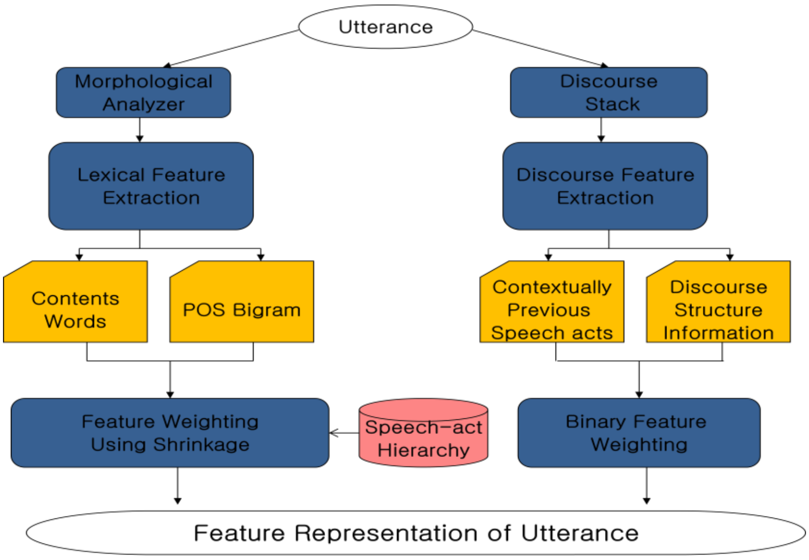
Utterance
Morphological Analyzer
Discourse Stack
Lexical Feature Extraction
Discourse Feature Extraction
Contents Words
POS Bigram
Contextually Previous Speech acts
Discourse Structure Information
Feature Weighting Using Shrinkage
Speech-act Hierarchy
Binary Feature Weighting
Feature Representation of Utterance

Fig. 1 Proposed feature-extraction process and weighting method.

3.1. *Lexical feature extraction and weighting method using a speech-act hierarchy*

Since an utterance is approximated based on its features, the extraction and weighting of features is a very important process for speech-act classification. In this section, we first explain how lexical features and POS bigrams are extracted. Next, we present how to reflect contextual information using a speech-act hierarchy on feature weighting. A speech-act hierarchy is constructed using adjacency pairs, which give us the opportunity to apply a type of contextual information to speech-act classification through feature weighting. Since adjacency pairs group utterances into two parts, such as requests and answers, according to their function, a large amount of information from speech acts with similar functions within a dialogue can provide complimentary clues for the speech-act classification as contextual information.

3.1.1 *Lexical feature extraction using a morphological analyzer*

Many previous studies on Korean speech-act classification have applied syntactic patterns as intra-utterance features. Although a syntactic pattern can represent the syntactic and semantic features of utterances, previous studies have found that syntactic patterns from a conventional syntactic parser are incomplete owing to errors in the syntactic analysis and are dependent on time-consuming, manually generated knowledge [18, 20]. To solve this problem, our lexical feature extraction method uses only the analysis results from a morphological analyzer so that our method becomes more robust to errors propagated from basic language analysis; we believe that a morphological analyzer generally generates fewer errors than a syntactic analyzer [15, 21].

We assume that content words and POS tag sequences in an utterance can provide very effective information for detecting the speech act of that utterance. Based on this assumption, we extract informative features for speech-act analysis using only a morphological analyzer. Lexical features include content words annotated with POS tags and POS bigrams of all words in an utterance. Content words generally have noun, verb, adjective and adverb POS tags. For example, the lexical features of utterance (5) in Table 1 consist of four content words and seven POS bigrams, as illustrated in Figure 2.

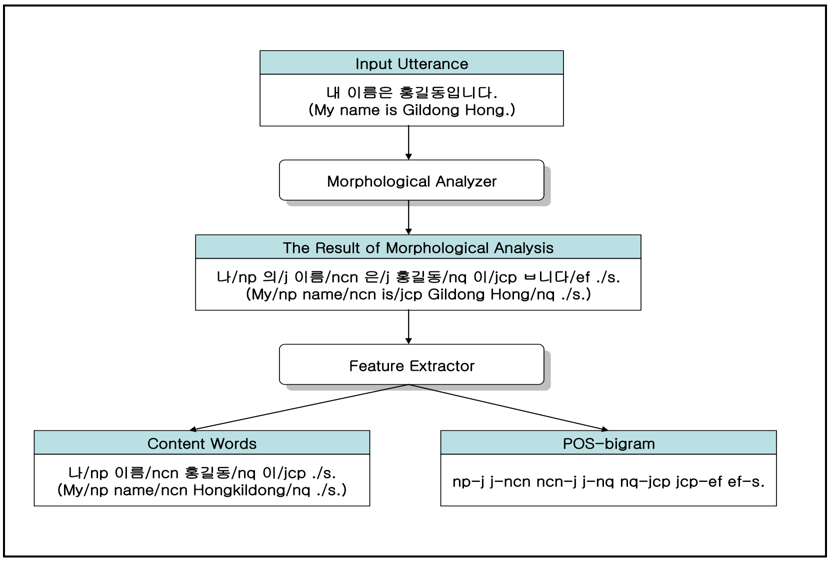
Input Utterance
내 이름은 홍길동입니다.
(My name is Gildong Hong.)
Morphological Analyzer
The Result of Morphological Analysis
나/np 의/j 이름/ncn 은/j 홍길동/nq 이/jcp ㅂ니다/ef ./s.
(My/np name/ncn is/jcp Gildong Hong/nq ./s.)
Feature Extractor
Content Words
나/np 이름/ncn 홍길동/nq 이/jcp ./s.
(My/np name/ncn Hongkildong/nq ./s.)
POS-bigram
np-j j-ncn ncn-j j-nq nq-jcp jcp-ef ef-s.

Fig. 2 Example of lexical feature extraction using a morphological analyzer.

3.1.2 *Lexical feature weighting scheme using a two-level speech-act hierarchy*

A speech-act hierarchy is constructed using pragmatic knowledge such as adjacency pairs. Utterances are often paired according to their function, such as a request and response pair. These adjacency pairs are defined as pairs of utterances that are adjacent and ordered as first and second parts; a particular type in the first part requires a particular type for the second part, such as ask-if/ask-ref/ask-confirm *vs*. response and offer/request/suggest *vs*. accept/reject. Many dialogues are also closed using the utterance of “Thank you.” according to the habitual characteristics of dialogues and idiomatic expressions of daily life. A two-level speech-act hierarchy constructed using these kinds of pragmatic knowledge are listed in Table 2. The first level of this hierarchy is composed of four different types of utterances (request, response, emotion and common use types) whereas the second level of each type contains appropriate speech acts.

Table 2. Two-level speech-act hierarchy

1st level	**Request Type**	**Response Type**	**Emotion Type**	**Common Use Type**
2nd level	Ask-if, Ask-ref, Ask-confirm, Offer, Suggest, Request	Accept, Response, Reject, Acknowledge	Expressive, Promise, Closing,	Opening, Correct, Inform, Introduce-oneself

This two-level speech-act hierarchy provides useful information for feature weighting. That is, the probability of a feature existing in a particular type including a speech act plays a complimentary role in the classification of the speech act. The degree of occurrence of features in a type within a dialogue can be considered as contextual information; the additional probabilistic information for similar speech acts in each type generally provides rich classification power to speech-act classifiers. Moreover, it can partially resolve the data sparseness problem that often occurs in a speech-act corpus since it is very difficult to collect and create sufficient numbers of dialogue examples tagged with large quantities of information for various areas of application. In particular, the low frequency of certain speech acts has created serious data sparseness problems in several previous studies; the accuracy of each speech act has tended to be proportional to the frequency with which each speech act occurs in the training data [16-18, 20]. Shrinkage is employed as a statistical technique to estimate a lexical feature weight by using the two-level speech-act hierarchy [22]. This shrinkage technique provides a very effective way to utilize a two-level speech-act hierarchy for applying contextual information and resolving the data sparseness problem. Finally, the two-level hierarchy shrinks parameter estimates in data-sparse speech acts of the second level toward estimates in a data-rich type of the first level in the most optimal way under given conditions.

The shrinkage technique estimates the probability of a feature as the weighted sum of the maximum-likelihood estimates from the leaf to root levels in a hierarchy [22]. Figure 3 illustrates how the shrinkage-based estimate of the probability of a feature ("나/np") in a given speech act ("ask-if") can be calculated using a weighted sum of the maximum-likelihood estimates from the leaf to root.

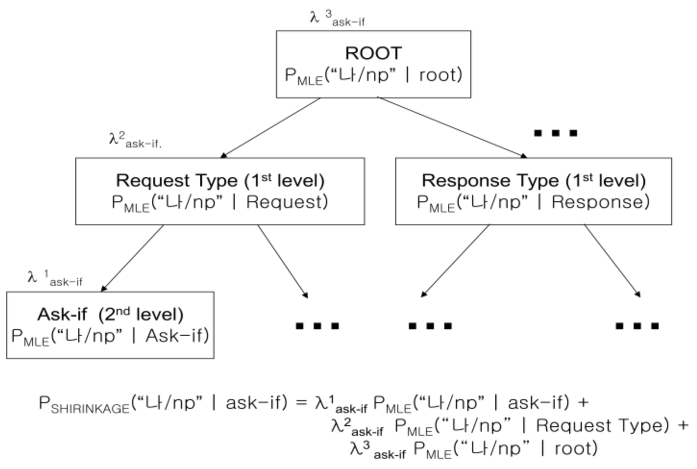

Fig. 3 Shrinkage-based estimation of feature probability.

Let $\{\hat{\theta}_j^1, \hat{\theta}_j^2, \hat{\theta}_j^3\}$ be three estimates of a speech act s_j in a hierarchy of speech acts, where $\hat{\theta}_j^1$ is the estimate at the second level (leaf), $\hat{\theta}_j^2$ at the first level, and $\hat{\theta}_j^3$ at the root level. Interpolation weights among the speech act s_j and its ancestors are written as $\{\lambda_j^1, \lambda_j^2, \lambda_j^3\}$, where $\sum_{i=1}^{3} \lambda_j^i = 1$. We can apply $\breve{\theta}_j$ for a new estimate of the conditioned feature probabilities of the speech act based on shrinkage. The new estimate for the probability of feature f_t given speech act s_j is as follows:

$$\breve{\theta}_{jt} = P(f_t \mid s_j; \breve{\theta}_j) = \lambda_j^1 \hat{\theta}_{jt}^1 + \lambda_j^2 \hat{\theta}_{jt}^2 + \lambda_j^3 \hat{\theta}_{jt}^3 . \tag{1}$$

Optimal interpolation weights were empirically derived using the following iterative procedure.

Initialization:

Set up each λ_j^i to certain initial values: $\lambda_j^i = \frac{1}{k}$

Iteration:

1. Calculate the degree to which each estimate predicts feature f_t in the held-out feature set, H_j, from s_j:

$$\beta_j^i = \sum_{f_t \in H_j} P(\hat{\theta}_j^i \text{ was used to generate } f_t) = \sum_{f_t \in H_j} \frac{\lambda_j^i \hat{\theta}_{jt}^i}{\sum_m \lambda_j^m \hat{\theta}_{jt}^m}$$

2. Compensate for the degree of loss caused by a large variation in each degree:

$$\beta_j^i = \beta_j^i + \frac{\sum_m \beta_j^m}{m}$$

3. Derive new weights by normalizing the β_j^i values:

$$\lambda_j^i = \frac{\beta_j^i}{\sum_m \beta_j^m}$$

Terminate: Upon convergence of the likelihood function

Table 3 presents the resulting mixture weights learned using this procedure.

Table 3. Example of interpolation weights learned using shrinkage-based estimation.

Speech act		Interpolation Weights		
1st level	**2nd level**	**Root (λ^3)**	**1st level (λ^2)**	**2nd level (λ^1)**
Request Type	Ask-if	0.237	0.242	0.52
	Ask-ref	0.282	0.295	0.422
	Ask-confirm	0.212	0.215	0.571
	Offer	0.207	0.209	0.583
	request	0.24	0.247	0.512
	Suggest	0.217	0.22	0.562
Response Type	Accept	0.214	0.218	0.566
	Response	0.367	0.329	0.302
	Reject	0.212	0.215	0.571
	Acknowledge	0.231	0.237	0.531
Emotion Type	Expressive	0.229	0.279	0.49
	Promise	0.227	0.28	0.491
	Closing	0.222	0.256	0.521
Common Use Type	Opening	0.23	0.251	0.517
	Introducing-oneself	0.233	0.249	0.517
	Correct	0.205	0.211	0.583
	Inform	0.26	0.332	0.406

3.2. *Discourse feature extraction method generated by a discourse stack and discourse rules*

Many previous studies have used the speech acts of previous utterances as discourse features. When a sub-dialogue occurs, the speech act of a current utterance is not related to that of the immediately prior utterance. In this case, the current utterance has to be related to the speech act of an utterance before the sub-dialogue. For example, the speech act of the seventh utterance in Table 4 (UID: 7) must be linked not to that of the sixth utterance (UID: 6) but to that of the second utterance (UID: 2). In our discourse feature extraction method, a discourse stack is designed based on discourse rules to find the correct previous utterance related to the current utterance. In addition, the discourse stack can provide discourse structure information regarding the sub-dialogue, i.e., its beginning and end. The discourse features from the proposed method consist of contextually previous speech act and structural information regarding the discourse using a discourse stack.

Table 4. Examples of discourse features.

UID	Utterance	Speech Acts (DSI)	Type 1	Type 2	Stack Speech acts (UID)
(1)	방을 하나 예약하고 싶은데요. (I would like to reserve a room.)	Inform (NULL)	DS	DS, NULL	Empty
(2)	어떤 방을 원하시죠? (What kind of room do you want?)	Ask-ref (NULL)	Inform	Inform, NULL	Ask-ref (2)
(3)	어떤 종류의 방이 있습니까? (What kind of rooms do you have?)	Ask-ref (SS)	Ask-ref	Ask-ref, NULL	Ask-ref (3) Ask-ref (2)
(4)	더블룸과 싱글룸이 있습니다. (We have single and double rooms.)	Response (SE)	Ask-ref	Ask-ref, SS	Ask-ref (2)
(5)	방값이 얼마죠? (How much are those rooms?)	Ask-ref (SS)	Response	Response, SE	Ask-ref (5) Ask-ref (2)
(6)	싱글은 삼만원이고 더블은 사만원 입니다. (Single rooms cost 30,000 won and double rooms cost 40,000 won.)	Response (SE)	Ask-ref	Ask-ref, SS	Ask-ref (2)
(7)	싱글룸으로 해주세요. (A single room, please.)	Response (NULL)	Response	Ask-ref, SE	Empty

* UID: ID of utterance, Type 1: using a speech act of the immediately prior utterance as a feature (discourse feature of previous studies), Type 2: using a discourse stack (the proposed discourse feature), Discourse stack information (DSI): dialogue-start (DS), sub-dialogue start (SS), sub-dialogue end (SE).

The discourse stack is a simple system that uses a stack procedure and the functional characteristic of utterances, such as adjacency pairs. According to the adjacency pairs, utterances are often paired according to their function, such as a request and response pair [23, 24]. This means that a particular type in the first part (request type) requires a matching type for the second part (response type). This constraint can be removed to allow more dependences between utterances. For example, remote links should be allowed between the first and second parts, since other utterances are sometimes inserted between them (e.g., a clarification sub-dialogue). This functional characteristic of utterances can be applied to a discourse stack. All speech acts are divided into three types according to their function in a dialogue: the first type is a request used as the first part of a dialogue, the second is a response used as the second part of a dialogue, and the third is a lone type that cannot be formed into an utterance pair. The request and response types are defined as in a speech-act hierarchy as listed in Table 5, but the lone type is defined as a union of emotion and common use types in a speech-act hierarchy.

Table 5. Three types of speech acts in a dialogue corpus.

Types of Speech Act	*Request Type*	*Response Type*	*Lone Type*
Speech Acts	ask-ref, ask-if, ask-confirm, offer, suggest, request	accept, response, reject, acknowledge	opening, introducing-oneself, correct, inform, expressive, promise, closing

We consider the dialogue state that other utterances are inserted in between an request-type utterance and a response-type one (adjacency pair) as dialogue segmentation. A sub-dialogue occurs when another utterance begins at this dialogue segmentation. The discourse stack is designed based on this theory of adjacent pairs and is implemented using the following algorithms.

```
# Discourse Structure Information (DSI): Sub-dialogue Start (SS), Sub-dialogue End (SE)

For each utterance
Begin
Discourse Feature Selection:
        If (Stack is Empty)
            Select the speech act and DSI of a previous utterance as a discourse feature
        Else
            Select the speech act at the top of the discourse stack and the DSI of the
            previous utterance
Operation:
        If (The speech act of the current utterance is a Request Type)
            If (Stack is not Empty)
                Assign SS to DSI of the current utterance
            Push speech act of the current utterance into the discourse stack
        Else If (The speech act of the current utterance is a Response Type)
            Pop the speech act into the discourse stack
            If (Stack is not Empty)
                Assign SE to the DSI of the current utterance
End
```

The discourse structure information (DSI) including sub-dialogue start (SS) and sub-dialogue end (SE) gives us additional information about whether or not the previous utterance occurs in any sub-dialogue. For example, utterances (3), (4), (6), and (7) of the dialogue in Table 4 have the same speech act ("ask-ref") as that of contextually previous utterance, but they are assigned to three distinguishable states according to the DSI of the contextually previous utterance:

“ask-ref, NULL,” “ask-ref, SS,” and “ask-ref, SE.” Note that an utterance with any request-type speech act (e.g., “ask-ref”) rarely occurs after an utterance tagged by another request-type speech act (e.g., “ask-ref”) and discourse structure information (“SS”) of a sub-dialogue start, because collaborative dialogues are commonly assumed for dialogue analysis.

Dialogues naturally generate exceptions that are not handled by the proposed algorithm. In particular, 3.5% of utterances in our dialogue corpus are exceptions. After we analyze such utterances, we can summarize the reasons and their solutions as listed in Table 6.

Table 6. Heuristic rules for exception types

Exception Types and Reasons	Solution by Heuristic Rule
1. Speaker asks two questions and listener answers each question.	If an request-type utterance has no answer after four exchanges, remove the pushed speech act of the utterance in the discourse stack.
2. Listener does not make an answer.	
3. Reaction is substituted for a response utterance corresponding to request-type utterances such as “suggest” and “offer.”	

3.3. *Embodying entire features into speech-act classifiers*

This section describes the composition of an entire feature set and explains how feature probabilities estimated through shrinkage can be applied to feature weights for speech-act classification. First, the entire feature set consists of lexical and discourse features as illustrated in Figure 4. The discourse features are composed of a speech act of contextually previous utterance from the discourse stack and discourse structure information including *DS*, *SS*, and *SE*.

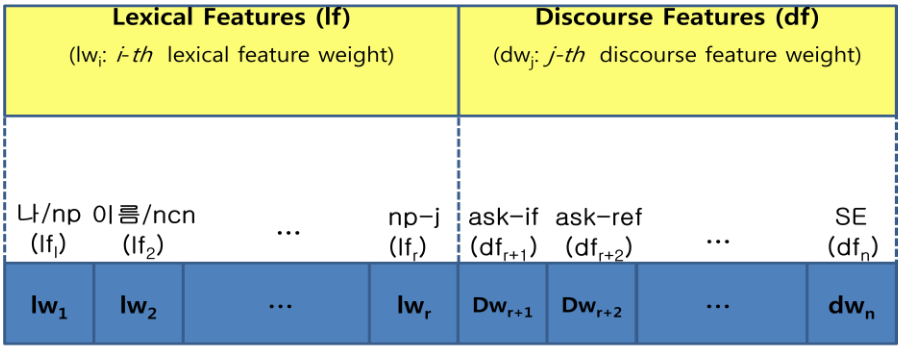

Fig. 4 Composition of an entire feature set (lexical and discourse features).

Estimated feature probabilities can be easily used in a probabilistic model such as the naive Bayes classifier; $P(f_t|s_j)$ in the naive Bayes formula can be replaced with $P(lf_t|s_j;\breve{\theta}_j)$ for lexical features and $P_{MLE}(df_k|s_j;\hat{\theta}_j^1)$ for discourse features in formula (1) as follows:

$$P(s_j \mid u_i) = \frac{P(s_j)\prod_{t=1}^{r} P(lf_{u_i,t} \mid s_j;\breve{\theta}_j) \prod_{k=r+1}^{n} P_{MLE}(df_{u_i,k} \mid s_j;\hat{\theta}_j^1)}{P(u_i)} . \tag{2}$$

where u_i is i-th utterance, s_j is i-th speech act, $lf_{u_i,t}$ is the t-th lexical feature occurred in the i-th utterance u_i, and $df_{u_i,k}$ is the k-th discourse feature occurred in the the i-th utterance u_i.

For vector-based models such as SVM, the various weighting scheme is applied [25] but the binary feature-weighting scheme performs well in speech-act classification because each feature in an utterance rarely occurs more than once [7]. In addition, the lexical feature probabilities estimated by shrinkage are applied to binary feature weighting as follows:

$$lw_t = \begin{cases} 0 & if\ \text{nonexistent} \\ 1.0 + P(lf_t|s_j;\breve{\theta}_j) & otherwise \end{cases} , \quad dw_t = \begin{cases} 0 & if\ \text{nonexistent} \\ 1.0 & otherwise \end{cases} . \tag{3}$$

4. Empirical Evaluation

4.1. *Experimental Data*

We used a Korean dialogue corpus, which was transcribed from real conversations such as those occurring when making hotel, airline, and tour reservations. This corpus consists of 528 dialogues and 10,285 utterances (19.48 utterances per dialogue) [2,15,17,18]. In total, 17 types of speech acts were used in this dialogue corpus. Table 8 lists their distribution. In the experiment, the Korean dialogue corpus was divided into training (428 dialogues and 8,349 utterances) and testing data (100 dialogues and 1,936 utterances).

Table 7. Speech-act distribution of the corpus.

Speech-Act Type	Distribution (%)	Speech-Act Type	Distribution (%)
Accept	2.49	Introducing-oneself	6.75
Acknowledge	5.75	Offer	0.4
Ask-confirm	3.16	Opening	6.58
Ask-if	5.36	Promise	2.42
Ask-ref	13.39	Reject	1.07
Closing	3.39	Request	4.96
Correct	0.03	Response	24.73
Expressive	5.64	Suggest	1.98
Inform	11.9	**Total**	**100**

We followed the standard definition of precision when measuring the performance. The precision values were first computed for each speech acts and averaged as a global measure, which is referred to as micro-averaging [26]. All of our experiments employed the naive Bayes (NB) and SVM classifiers, which are representative learning models for probability and vector models, respectively.

4.2. *Experimental Results*

A speech-act classification system with only lexical features was used as a baseline in our experiments for comparison with the proposed methods. In addition, we compared the baseline system with an existing method that uses syntactic patterns. The performances of each classifier are listed in Table 8.

Table 8. Performance of the baseline system comparing with an method using syntactic patterns.

	NB	SVM
Syntactic Pattern Feature	61.05	68.33
Only Lexical Feature (Baseline System)	71.22	79.95

As can be seen in Table 8, the speech-act classification system using only lexical feature performed much better than the system using syntactic patterns in both classifiers; NB improved by 16.65% and SVM by 17.01% over the system using syntactic patterns. These experimental results indicate that our lexical-feature extraction method is very effective at extracting features for speech-act classification.

4.2.1 *Verifying the proposed system using lexical and discourse features reflecting contextual information*

In this section, the discourse features are first verified. The contextually previous speech-act and discourse structure information (sub-dialogue start (SS) and sub-dialogue end (SE)) from the discourse stack are added to the entire feature set. Moreover, we conducted an additional experiment to verify the proposed discourse features more precisely; the entire features of the additional experiment consist of lexical features plus the speech-act of immediately prior utterance, instead of our proposed discourse features. Table 9 presents the experimental results and Figure 5 shows their comparison.

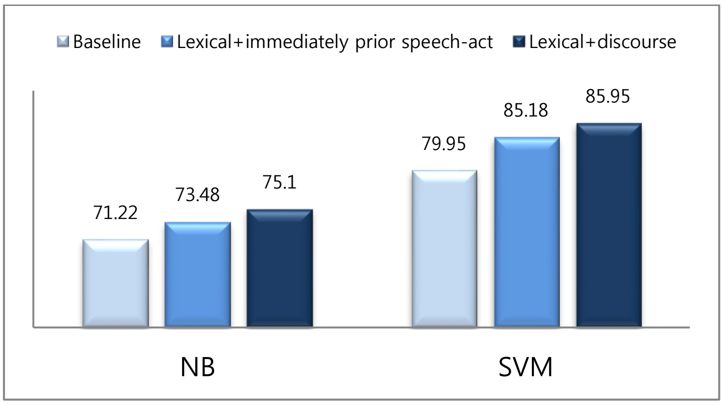

Fig. 5 Comparison of performances of each method for each classifier.

Table 9. Performances of the proposed method using discourse features.

	NB	SVM
Baseline System	71.22	79.95
Lexical Feature + Speech-act Feature of Immediately Prior Utterance	73.48	85.18
Lexical Feature + Discourse Features from Discourse Stack	**75.1**	**85.95**

As listed in Table 9, the discourse features from the proposed discourse stack are more effective as contextual information than the speech-act feature of the immediately prior utterance in both classifiers. In particular, the performances of the classifiers using the proposed discourse features achieved 3.88% and 6% improvements in comparison to those of the baseline system on NB and SVM, respectively.

Next, we verified the proposed feature-weighting method using a two-level speech-act hierarchy and the shrinkage technique as further contextual information. As shown in Table 10, the proposed feature-weighting method achieved greater improvement in performance than the model without two-level speech-act hierarchy for both classifiers. Finally, the proposed model achieved improvements of 6.18% (NB) and 6.65% (SVM) over the baseline system.

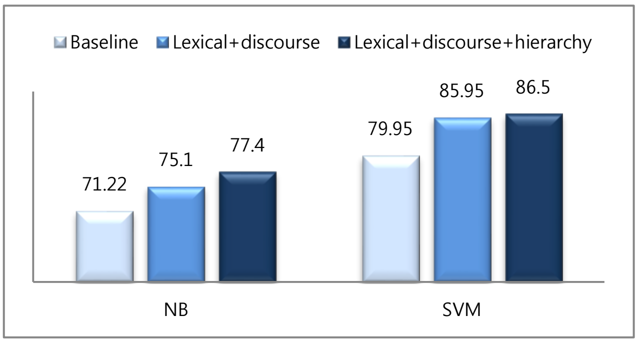

Fig. 6 Comparison of the performances of each method for each classifier.

Table 10. Performances of the proposed feature-weighting method using a two-level speech-act hierarchy.

	NB	SVM
Baseline System	71.22	79.95
Lexical Feature + Discourse Features from Discourse Stack	75.1	85.95
Lexical Feature + Discourse Features from Discourse Stack + Two-level Speech-act Hierarchy	**77.4**	**86.5**

Since the proposed model effectively applies contextual information to each speech-act classification, it can improve the performance of speech-act classifiers, particularly on speech acts with a small number of utterances such as “*accept*,” “*closing*,” “*offer*,” “*reject*,” and “*suggest*.” This strong aspect of the proposed model can be observed in Figure 7 and Table 11, which show the results of an experiment conducted using SVM as a classifier and macro-averaging as a global measure over all speech acts.

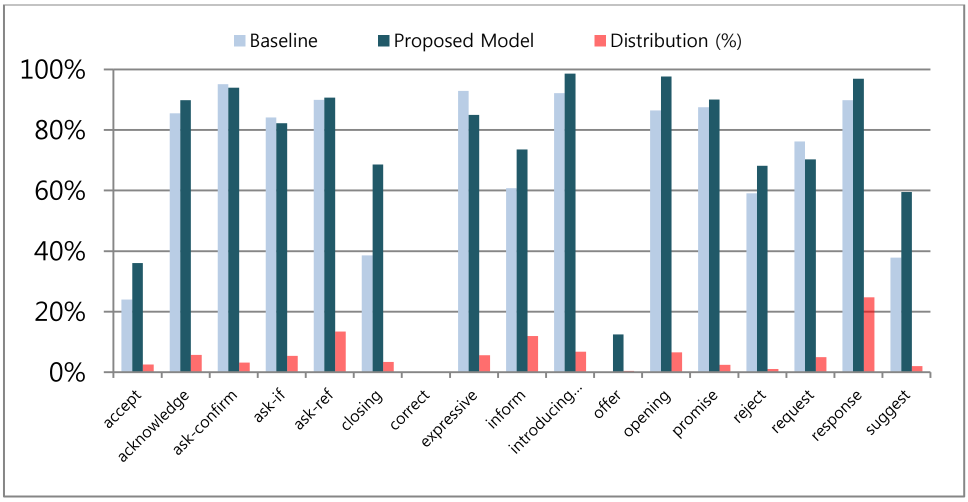

Fig. 7 Comparison of performances between the baseline system and the proposed model according to the distribution of each speech act.

Table 11. Performance comparison between the baseline system and the proposed model according to the distribution of each speech act.

Speech-Act Type	**Baseline System**	**Proposed Model**	**Distribution (%)**	**Speech-Act Type**	**Baseline System**	**Proposed Model**	**Distribution (%)**
Accept	24.0	36.0	2.49	Introducing -oneself	92.2	98.58	6.75
Acknowledge	85.51	89.86	5.75	Offer	0.0	12.5	0.4
Ask-confirm	95.12	93.9	3.16	Opening	86.4	97.6	6.58
Ask-if	84.16	82.18	5.36	Promise	87.5	90.0	2.42
Ask-ref	89.88	90.66	13.39	Reject	59.09	68.18	1.07
Closing	38.57	68.57	3.39	Request	76.19	70.24	4.96
Correct	0.0	0.0	0.03	Response	89.88	96.9	24.73
Expressive	92.92	84.96	5.64	Suggest	37.84	59.46	1.98
Inform	60.80	73.6	11.9	***macro-averaging***	***64.7***	***71.37***	

4.2.2 *Comparison of the proposed model with other speech-act analysis models*

We compare the proposed model with two previous speech-act analysis models that used the same experimental data: Choi's Model (**CHOI**) [17] and Lee's Model (**LEE**) [18]. Table 12 shows the summary of techniques used each speech-act analysis models.

Table 12. Summary of Techniques used in the proposed and previous speech-act analysis models

	Linguistic Information	***Sub-dialog Processing***	***Data Sparseness processing***
CHOI	morphological features + syntactic patterns	statistical model using previous utterances	N/A
LEE	morphological features + syntactic patterns	previous utterances	N/A
Proposed Model	morphological features	discourse stack using adjacency pairs	shrinkage technique using speech-act hierarchy

Table 13 lists the results from each speech-act analysis models. As a result, our proposed model yielded the best result among the three, 4.6% better than others.

Table 13. Experimental results of the proposed and previous models.

Model	**Precision (%)**
CHOI	81.9
LEE	81.5
Proposed Model	**86.5**

We think that this improvement is caused by effectively reflecting contextual information into speech-act analysis using a discourse stack and the shrinkage technique. Especially, they show the effectiveness to resolve the problems from the lack of sub-dialogue consideration and the data sparseness problem. In addition, the proposed system is free from some problems of syntactic patterns incompleteness because it does not require any syntactic information.

5. Conclusions

This paper has presented an effective speech-act classification model to utilize contextual information. The proposed model uses a new feature-weighting scheme using a two-level hierarchy and the shrinkage technique, and an effective discourse-feature extraction scheme using adjacency pairs. They both were experimentally verified as very effective methods in speech-act classification. Finally, the proposed model achieved about over 6% improvements in both classifiers when they were compared to the baseline model. In particular, the proposed model achieved a high improvement for sparse speech-act classes. In addition, the proposed model showed better performance than other previous speech-act classification models in our experiments.

References

[1] B. Grosz, "Discourse and dialogue," In *Survey of the State of the Art in Human Language Technology*, Center for Spoken Language Understanding, (1995), pp. 227-254.

[2] J. Lee, G. Kim, and J. Seo, "A dialogue analysis model with statistical speech-act processing for dialogue machine translation," In *Proc. of the ACL Workshop on Spoken Language Translation*, (1997), pp. 10-15.

[3] S. Rangarajan, S. Narayanan, and S. Bangalore, "Modeling the intonation of discourse segments for improved online dialog act tagging," In *Proc. of ICASSP 2009*, (2009), pp. 5033-5036.

[4] A. Stolcke, K. Ries, N. Coccaro, E, Shriberg, R. Rates, D. Jurafsky, P. Taylor, C. Van EssDykema, R. Martin, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, Vol. 26, No. 3, (2000), pp. 339-373.

[5] A. Clark and A. Popescu-Belis, "Multi-level dialogue act tags," In *Proc. of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, (2004), pp. 163-170.

[6] D. R. Traum, “20 questions for dialogue act taxonomies,” *Journal of Semantics*, Vol. 17, No. 1,(2000), pp. 7-30.

[7] A. Dielmann and S. Renals, "Recognition of dialogue acts in multiparty meetings using a switching DBN," *IEEE Transaction on Audio, Speech, and Language Processing*, Vol. 16, No. 7, (2008), pp. 1303-1314.

[8] S. Rangarajan, S. Narayanan, and S. Bangalore, "Modeling the intonation of discourse segments for improved online dialog act tagging," In *Proc. of ICASSP 2009*, (2009), pp. 5033-5036.

[9] K. Laskowski and E. Shriberg, "Modeling other talkers for improved dialog act recognition in meetings," *Proceedings of INTER-SPEECH 2009*, (2009), pp. 2783-2786.

[10] K. Laskowski and E. Shriberg, "Comparing the contributions of context and prosody in text-independent dialog act recognition," In *Proc. of ICASSP 2010*, (2010), Dallas TX, USA.

[11] A. Stolcke, N. Coccaro, R. Bates, P. Taylor, and C. V. Ess-Dykema, “Dialogue act modeling for automatic tagging and recognition of conversational speech,” *Computational Linguistics*, Vol. 26, No. 3,(2000) pp. 1-3.

[12] S. Caberry, "A pragmatics-based approach to ellipsis resolution," *Computational Linguistics*, Vol. 15, No. 2, (1989), pp. 75-96.

[13] L. Lambert, *Recognizing Complex Discourse Acts: A Tripartite Plan-based Model of Dialogue*, Ph.D. Dissertation, Newark, Delaware: University of Delaware, (1993).

[14] H. Lee, *Analysis of Speech-acts for Korean Dialog Sentences*, MS Thesis, Sogang University, South Korea (1996).

[15] K. Kim, H. Kim, and J. Seo, "A neural network model with feature selection for Korean speech-act classification," *International Journal of Neural Systems*, Vol. 14, No. 6, (2004), pp. 407-414.

[16] K. Samuel, S. Caberry, and K. Vijay-Shanker, "Automatically selecting useful phrases for dialogue act tagging," In *Proc. of the Fourth Conference of the Pacific Association for Computational Linguistics*, Waterloo, Ontario, Canada, (1999).

[17] W. Choi, J. Cho, and J. Seo, "Analysis system of speech-acts and discourse structures using maximum entropy model," In *Proc. of COLING-ACL 99*, (1999), pp. 230-237.

[18] S. Lee and J. Seo, "A Korean speech-act analysis system using hidden Markov model with decision trees," *International Journal of Computer Processing of Oriental Languages*, Vol. 15, No. 3, (2002), pp. 231-243.

[19] N. Reithinger and M. Klesen, "Dialogue act classification using language models," In *Proc. of EuroSpeech 97*, Rhodes, Greece, (1997), pp. 2235-2238.

[20] N. Webb, M. Hepple, and Y. Wilks, "Error analysis of dialogue act classification," In *Proc. of the 8th International Conference on Text, Speech and Dialogue*, Carlsbad, Czech Republic, (2005).

[21] C. Yuan, X. Wang and F. Ren "Exploiting lexical information for function tag labeling," *International Journal of Innovative Computing, Information and Control (IJICIC)*, Vol. 6, NO. 3(B), (2010), pp. 1471-1480.

[22] A. MacCallum, R. Rosenfeld, T. Mitchell, and A. Y. Ng, "Improving text classification by shrinkage in a hierarchy of classes," In *Proc. of the International Conference on Machine Learning*, (1998).

[23] E. Schegloff and H. Sacks, "Opening up closing," *Semiotica*, Vol. 7, No. 4, (1973), pp. 289-327.

[24] S. Levinson, *Pragmatics, Cambridge*, UK: Cambridge University Press, (1983).

[25] R. Chen, and S. Chen, "Intrusion detection using a hybrid support vector machine based on entropy and TF-IDF," *International Journal of Innovative Computing, Information and Control (IJICIC)*, Vol. 4, NO. 2, (2008), pp. 413-424.

[26] Y. Yang, "An evaluation of statistical approaches to text categorization," *Information Retrieval Journal*, Vol. 1, Nos. 1-2, (1999), pp. 67-88.