JCSE Journal of Computing Science and Engineering

Regular Paper

Journal of Computing Science and Engineering,
Vol. 5, No. 4, December 2011, pp. 288-293

Review of Korean Speech Act Classification: Machine Learning Methods

Harksoo Kim
Department of Computer and Communications Engineering, Kangwon National University, Chuncheon, Korea **nlpdrkim@kangwon.ac.kr**

Choong-Nyoung Seon
Department of Computer Science and Engineering, Sogang University, Seoul, Korea **wilowisp@sogang.ac.kr**

Jungyun Seo*
Department of Computer Science and Engineering/Interdisciplinary Program of Integrated Biotechnology, Sogang University, Seoul, Korea **seojy@sogang.ac.kr**

Abstract
To resolve ambiguities in speech act classification, various machine learning models have been proposed over the past 10 years. In this paper, we review these machine learning models and present the results of experimental comparison of three representative models, namely the decision tree, the support vector machine (SVM), and the maximum entropy model (MEM). In experiments with a goal-oriented dialogue corpus in the schedule management domain, we found that the MEM has lighter hardware requirements, whereas the SVM has better performance characteristics.

Category: Human computing

Keywords: Korean speech act classification; Machine learning method

I. INTRODUCTION

Goal-oriented dialogues such as appointment scheduling, call routing, and hotel reservation booking consist of sequences of goal-oriented utterances. The speakers' intentions implied by each utterance can be represented using semantic forms called speech acts [1]. In Table 1, Utterance (3) shows that a user requests a system to search his schedule. The requesting action comprising Utterance (3) is the speech act.

As shown in Table 1, to generate correct reactions, a dialogue system should identify the speech acts indicated by users' utterances to capture the speaker's intentions. If a dialogue system fails to capture users' intentions, the system will not be able to decide whether to respond to users' questions or to request additional information from users to achieve the task goals. It is difficult, however, to infer speech acts from surface utterances because they are context-dependent. For example, the speech

Table. 1. Example of a goal-oriented dialogue annotated with speech acts

No.	Speaker	Utterance	Speech act
1	User	Hello.	Greeting
2	System	May I help you?	Opening
3	User	Tell me tomorrow's schedule.	Request
4	System	You have an appointment with Kildong Hong at 11 a.m.	Response
5	User	We changed the appointment.	Inform
6	System	What changed?	Ask_ref
7	User	The appointment date changed.	Response

Open Access http://dx.doi.org/10.5626/JCSE.2011.5.4.288 http://jcse.kiise.org

Received 17 June 2011, **Revised** 31 October 2011, **Accepted** 8 November 2011
*Corresponding Author

 pISSN: 1976-4677 eISSN: 2093-8020

act of Utterance (7) in Table 1 can be evaluated as an "inform" or "response" in surface analysis. Various models have been proposed over the past 25 years to resolve this ambiguity. In recent years, there has been increased interest in using statistical and machine learning approaches. In this paper, we review a representative probabilistic model for speech act classification and then compare various machine learning models to it.

This paper is organized as follows. In Section II, we review earlier work on speech act identification. In Section III, we review speech act classification models based on machine learning methods. In Section IV, we experimentally compare the reviewed models. And finally, in Section V, we draw conclusions.

II. EARLIER WORK

Initial approaches to speech act identification have been based on knowledge such as plan inference recipes and domain-specific knowledge [2, 3]. Since these knowledge-based models depend on costly handcrafted knowledge, they are difficult to scale up and expand to other domains. To overcome this problem, in recent years, many machine learning approaches have been proposed for speech processing. The task of identifying users' intentions is an example of an area in which this approach has been relatively successful as shown using various machine learning models [4-14]. Machine learning models offer a way to associate utterance features with particular categories indicating users' intentions since the computer can efficiently analyze a large quantity of data and consider many different feature interactions. However, input features critically affect machine learning models. If the input features are uninformative and biased, they do not take full advantage of particular input features that may provide valuable clues in identifying users' intentions. Many feature extraction and selection methods have been proposed to resolve these problems. Lee et al. [5] have used the structural information of discourse in speech act analysis. However, such structural information is insufficient for covering various dialogues since the authors used a restricted rule-based model such as a recursive transition network to perform the discourse structure analysis. Kim et al. [15] comparatively studied optimal feature identification for Korean speech act classification. They evaluated and compared each feature combination. Many researchers have studied feature selection methods for text categorization and speech act classification [16, 17]. Yang and Pedersen [17] present a comparative study of the feature selection methods for statistical learning in text categorization. They found information gain and use of the χ^2 statistic to be most effective in the experiments. Kim et al. [16] proposed that a neural network can partially increase precision and decrease training time using the feature selection method based on the χ^2 statistic.

III. SPEECH ACT CLASSIFICATION MODEL

A. Representative Probabilistic Model for Speech Act Classification

Given n utterances $U_{1,n}$ in a dialogue, let $S_{1,n}$ denote the speech acts of $U_{1,n}$. The speech act classification model can then be formally defined as follows:

$$SA(U_{1,n}) \overset{def}{=} \arg\max_{S_{1,n}} P(S_{1,n}|U_{1,n})\,. \tag{1}$$

We can rewrite Equation (1) as Equation (2) using the Bayes theorem. We exclude $P(U_{1,n})$ from Equation (s) since it is always constant for $S_{1,n}$:

$$\begin{aligned} SA(U_{1,n}) &\overset{def}{=} \arg\max_{S_{1,n}} \frac{P(U_{1,n}|S_{1,n})P(S_{1,n})}{P(U_{1,n})} \\ &\overset{drf}{\approx} \arg\max_{S_{1,n}} P(U_{1,n}|S_{1,n})P(S_{1,n})\,. \end{aligned} \tag{2}$$

Next we simplify Equation (2) by making the following assumptions: the current speech act is only dependent on earlier speech acts, and the current utterance is dependent on its speech act. With these assumptions, we formulate the speech act classification model as a product of sentential probability $P(S_i|U_i)$ and contextual probability $P(S_i|S_{1,i-1})$ [4, 18] as shown in Equation (3):

$$SA(U_{1,n}) \approx \arg\max_{S_{1,n}} \prod_{i=1}^{n} P(S_i|U_i)P(S_i|S_{1,i-1})\,. \tag{3}$$

Equation (3) is a representative probability model (a so-called hidden Markov model, HMM), which has been the basis of many machine learning approaches to speech act classification.

B. Machine Learning Models Adopted in Speech Act Classification

The earlier machine learning approaches for speech act classification can be divided into 3 groups: rule-, margin-, and statistics-based. The main idea of the rule-based group is to automatically generate a set of ordered rules from a training corpus. Transformation-based learning (TBL) and a decision tree (DT) are often used for speech act classification [7, 19]. The central idea of TBL is to learn an ordered set of symbolic rules according to their contribution to the training corpus. A DT is a decision-making mechanism that automatically generates possible choices according to information gain. TBL and the DT offer the advantage of having human-interpretable rules that can be manually edited for performance tuning.

The main idea of the margin-based group is to identify the most effective decision boundaries that separate positive examples and negative examples in a vector space. A support vector machine (SVM) and a multilayer perceptron (MLP; a feed forward artificial neural network model) have shown good performance in speech act classification [20, 21]. The goal of an SVM is to find the particular hyperplane that maximizes the margin of separation between a cluster of positive examples and a cluster of negative examples. An SVM transforms the given non-linear problems into linear problems by projecting input features into higher dimensions and then quickly solving the given problems high performance. An SVM is one of the best known binary classification models. The goal of MLP is to find the set of weight values that will cause the neural network output to match the actual target values as closely as possible. In particular, anything that can be represented as a mapping between vector spaces can be approximated to arbitrary precision by MLP (the most frequently used type). In practice, MLP is especially useful for solving mapping problems to which hard and fast rules cannot easily be applied.

The goal of the statistics-based group is to overcome the following weak points of an HMM: the observation bias problem and the label bias problem. A maximum entropy model (MEM) and conditional random fields (CRFs) are representative statistical models that are adopted in speech act classification [13, 22]. A MEM focuses on relaxing the 2 independence assumptions of the HMM mentioned in Section III-A. Due to the strong independence assumptions, the observation targets of the HMM are restricted to atomic entities such as words and parts of speech (POS). In particular, it is not practical to represent multiple interacting features or long-range dependencies of the observations [23]. In Equation (3), all terms of the right hand side are represented by conditional probabilities. We can estimate the probability of each term using Equation (4):

$$P(a|b) = \frac{P(a,\, b)}{\sum_{a'} P(a',\, b)}. \tag{4}$$

Now we can evaluate $P(a,b)$ using the MEM shown in Equation (5) [24]:

$$P(a,\, b) = \pi \prod_{i=1}^{k} \alpha_i^{f_i(a,b)}, \text{ where } 0 < \alpha_i < \infty,\ i = \{1, 2, ..., k\}. \tag{5}$$

In Equation (5), a is a speech act, depending on the term, b is the context of a, π is a normalization constant, while α_i is the model parameter corresponding to each feature function, f_i. CRFs are focused on resolving the problem of transition probabilities being locally normalized (the so-called label bias problem): the transitions leaving a given state compete only against each other rather than against all of the other transitions in the model [23] as shown in Equation (6):

$$P(a|b) = \frac{P(a,\, b)}{\sum_{a'} P(a',\, \bar{o})}, \text{ where } \bar{o} \text{ is the entire obervation sequence.} \tag{6}$$

Machine learning model performance is critically affected by the quality of the input features (*i.e.*, how informative the input features are). Therefore, many researchers have performed various feature extraction methods [5, 15, 16]. Kim et al. [15] comparatively studied optimal feature identification for Korean speech act classification. Table 2 shows a set of optimal features proposed by Kim et al. [15].

As shown in Table 2, input features for speech act classification are divided into two types: one pertains to the input features associated with the sentential probability $P(S_i|U_i)$ in Equation (3), while the other pertains to the input features associated with the contextual probability $P(S_i|S_{1,i-1})$ in Equation (3). The former are generally called sentential features, while the latter are called contextual features. In many cases, sentential features are too numerous to be used as inputs to machine learning models. Therefore, methods of removing non-informative features have been required. Yang and Pedersen [17] performed a comparative study of optimal feature selection for document classification. They showed that the χ^2 statistic outperforms mutual information and information gain in document classification. The χ^2 statistic measures the lack of independence between a feature, f, and a category, S (*i.e.*, a speech act) as shown in Equation (7):

Table 2. Optimal feature set for Korean speech act classification

Type of features	Optimal features
N-gram	Morpheme-parts of speech pair
Last predicate information	Last word
	Last verb
	Last adverb
	The endings of a word
Grammatical morpheme sequence/set	Grammatical morpheme sequence
Surface information	Length of utterance (S/M/L)
Context information	The previous speech act of a partner's utterance
	The previous speech act of a speaker's utterance

$$\chi^2(f,\, S) = \frac{(A+B+C+D)\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)}. \tag{7}$$

In Equation (7), A is the number of times that f and S co-occur, B is the number of times that f occurs without S, C is the number of times that S occurs without f, and D is the number of times neither S nor f occur. To remove non-informative features, the maximum χ^2 statistic of a feature-category pair is calculated, as shown in Equation (8), and the top-n features are selected according to the feature scores:

$$\chi^2_{\max}(f) = \max_{k=1}^{m} \{\chi^2(f,\, S^k). \tag{8}$$

In Equation (8), S^k is the k^{th} instance among m speech acts.

VI. EXPERIMENTS

A. Data Sets and Experimental Settings

We collected a Korean dialogue corpus simulated in a schedule management domain similar to appointment scheduling and alarm setting. The dialogue corpus was obtained by eliminating interjections and erroneous expressions from the original transcriptions of simulated dialogues between two speakers, to whom a task of the dialogue had been given in advance: one participant freely asks something about his/her daily schedules, and the other participant responds to the questions or asks some questions in return, using knowledge bases provided in advance. This corpus consists of 900 dialogues, 20,079 utterances (22.3 utterances per dialogue). Each utterance in the dialogues is manually annotated with speech acts and concept sequences. Table 3 shows part of the annotated dialogue corpus.

In Table 3, KS represents a Korean sentence and EN represents the translated English sentence that is not written in the original dialogue corpus. SP has a value of either User or System depending on the speaker. SA represents a speech act. In this paper, we define 11 domain-independent speech acts (Table 4).

The manual tagging of speech acts was performed by five

Table 3. Part of the annotated dialogue corpus

Tag	Values
/ID/	3-9
/SP/	User
/KS/	약속 날짜와 장소가 바뀌었어.
/EN/	The appointment date and place were changed.
/SA/	Inform
/ID/	3-10
/SP/	System
/KS/	바뀐 날짜가 언제인가요?
/EN/	When is the changed date?
/SA/	Ask_ref
/ID/	3-11
/SP/	User
/KS/	12월 5일
/EN/	December 5
/SA/	Response

SP: a value of either User or System depending on the speaker, KS: Korean sentence, EN: the translated English sentence that is not written in the original dialogue corpus, SA: speech act.

graduate students with dialogue analysis knowledge and post-processed by a student in a doctoral course for consistency. To evaluate various machine learning models, we divided the annotated dialogue corpus into the training corpus (800 dialogues) and the testing corpus (100 dialogues). We selected a representative model per machine learning group for use as comparison models: a DT in the rule-based group, an SVM in the margin-based group, and a MEM in the statistics-based group. We selected MEM instead of CRFs in the statistics-based group because CRFs showed performance similar to the MEM despite the requirement of much more training time. We think that CRFs are more appropriate for batch jobs, such as POS tagging and named entity (NE) tagging, which are started after all strings have been input. The comparison models used the same input features as in Table 2. The numbers of features for machine learning methods are determined experimentally. A total of 3,000 sentential features were selected based on the χ^2 statistic in Equation (8) for each SVM and MEM. Because the feature selection did not improve DT performance, it used all of the sentential features (10,082 features). The toolkits used for implementations included C4.5 [25] for the DT, SVMlight [26] for the SVM, and MEMT [27] for the MEM. We set all parameters of each toolkit to default values.

B. Experimental Results

The first experiment performed evaluated the memory requirements and processing speeds of the various models. Table 5 shows the results of the first experiment. The comparison

Table 4. Speech acts and their meanings

Speech act	Description	Occurrence ratio in corpus
Greeting	The opening greeting of a dialogue	9.48
Expressive	The closing greeting of a dialogue	8.80
Opening	Sentences for opening a goal-oriented dialogue	0.02
Ask_ref	Wh-questions	22.52
Ask_if	Yn-questions	2.70
Response	Responses of Ask_ref, Ask_if, Request	37.99
Request	Declarative sentences for requesting actions	14.54
Ask_confirm	Questions for confirming previous actions	0.03
Confirm	Reponses of Ask_confirm	0.03
Inform	Declarative sentences for giving information	2.05
Accept	Agreement	1.83

Table 5. Comparison of memory requirements and processing speeds

Model	Training		Testing	
	Memory usage (MB)	Spending time (sec)	Memory usage (MB)	Response time (sec/utterance)
C4.5	276.69	1973.86	19.98	0.06
SVM	10.61	715.27	0.83	0.04
MEM	5.60	46.47	0.75	0.01

SVM: support vector machine, MEM: maximum entropy model.

Table 6. Performance comparison in terms of various evaluation measures

Model	Accuracy (%)	Macro precision (%)	Macro recall rate (%)	Macro F1-measure (%)
C4.5	91.64	85.78	86.25	86.01
SVM	93.38	87.83	89.60	88.71
MEM	92.76	86.70	88.02	87.36

SVM: support vector machine, MEM: maximum entropy model.

models were evaluated on a personal computer with an Intel Xeon 2.00 GHz CPU, 4 GB MB memory, and Red Hat Linux.

As shown in Table 5, the memory usage and spending time of a DT showed low performance compared to those of the other methods because the feature selection did not work in the DT. Because an SVM is a binary classification method, extension of the binary classification using an SVM is normally applied to *n*-ary classification. Therefore, the SVM requires more memory and computation time than does a MEM.

The second experiment compared the performance of the various models. Table 6 shows the model performance in terms of various evaluation measures such as the accuracy, macro precision, macro recall rate, and macro F1 measure.

In Table 6, the accuracy is the proportion of correct speech acts of those returned. The macro precision is the average proportion of correct speech acts per category of those returned. The macro recall rate is the average proportion of correctly returned speech acts per category of those that are correct. The macro F1-measure combines the macro precision and macro recall rate with an equal weighting in the following form: F1 = (2.0 × macro precision × macro recall rate)/(macro precision + macro recall rate). As shown in Table 6, an SVM shows the best performance, which is similar to Kim et al. [15], which reported that the MEM is also an efficient method for speech act classification because it has advantages in terms of hardware requirements and exhibits a performance of $<1\%$ compared with the SVM.

V. CONCLUSION

We reviewed the earlier machine learning methods for Korean speech act classification. First we reviewed a representative statistical model. Based on the statistical model, we reviewed three groups of machine learning models: a rule-based group, a margin-based group, and a statistics-based group. In the experiments with a goal-oriented dialogue corpus in a schedule management domain, we selected a single representative per group among previous models: C4.5 in the rule-based group, SVM in the margin-based group, and MEM in the statistics-based group. We then compared the representative models using various evaluation measures. The experimental results revealed that the MEM offers advantages in terms of hardware requirements while the SVM offers advantages in terms of performance.

ACKNOWLEDGEMENTS

This research was supported by Mid-career Researcher Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (MEST) (No.2011-0027537)/Basic Science Research Program through NRF grant funded by the MEST (No. 2010-0009875).

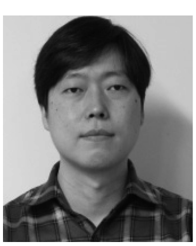
REFERENCES

1. J. Allen, Natural Language Understanding, Menlo Park, CA: Benjamin/Cummings, 1995.
2. L. Lambert and S. Carberry, “A tripartite plan-based model of dialogue,” *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics*, Berkeley, CA, 1991, June 18-21, pp. 47-54.
3. D. J. Litman and J. F. Allen, “A plan recognition model for subdialogues in conversations,” *Cognitive Science*, vol. 11, no. 2, pp. 163-200, 1987.
4. M. Nagata and T. Morimoto, “First steps towards statistical modeling of dialogue to predict the speech act type of the next utterance,” *Speech Communication*, vol. 15, no. 3-4, pp. 193-203, 1994.
5. J. Lee, G. C. Kim, and J. Seo, “A dialogue analysis model with statistical speech act processing for dialogue machine translation,” *Proceedings of Spoken Language Translation Workshop in conjunction with ACL/EACL*, Madrid, Spain, 1997, July 11, pp. 10-15.
6. N. Reithinger and M. Klesen, “Dialogue act classification using language models,” *Proceedings of EuroSpeech*, Rhodos, Greece, 1997, pp. 2235-2238.
7. K. Samuel, S. Carberry, and K. Vijay-Shanker, “Dialogue act tagging with transformation-based learning,” *Proceedings of the 17th International Conference on Computational Linguistics*, Montreal, QC, 1998, pp. 1150-1156.
8. A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. Van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, and M. Meteer, “Dialogue act modeling for automatic tagging and recognition of conversational speech,” *Computational Linguistics*, vol. 26, no. 3, pp. 339-373, 2000.
9. C. T. Langley, “Analysis for speech translation using grammar-based parsing and automatic classification,” *Proceedings of Student Research Workshop at the 40th Annual Meeting of the Association of Computational Linguistics*, Philadelphia, PA, 2002, July.
10. H. Kim and J. Seo, “An efficient trigram model for speech act analysis in small training corpus,” *Journal of Cognitive Science*, vol. 4, no. 1, pp. 107-120, 2003.
11. N. Webb, M. Hepple, and Y. Wilks, “Dialogue act classification based on intra-utterance features,” *Proceedings of the AAAI Workshop on Spoken Language Understanding*, Pittsburgh, PA, 2005, July 9-10.
12. W. S. Choi, H. Kim, and J. Seo, “An integrated dialogue analysis model for determining speech acts and discourse structures,” *IEICE Transactions on Information and Systems*, vol. E88-D, no. 1, pp. 150-157, 2005.
13. H. Lee, H. Kim, and J. Seo, “Domain action classification using

a maximum entropy model in a schedule management domain," *AI Communications*, vol. 21, no. 4, pp. 221-229, 2008.
14. S. Kang, H. Kim, and J. Seo, "A reliable multidomain model for speech act classification," *Pattern Recognition Letters*, vol. 31, no. 1, pp. 71-74, 2010.
15. M. J. Kim, J. H. Park, S. B. Kim, H. C. Rim, and D. G. Lee, "A comparative study on optimal feature identification and combination for korean dialogue act classification," *Journal of Korean Institute of Information Scientists and Engineers: Software and Applications*, vol. 35, no. 11, pp. 681-691, Nov 2008.
16. 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, pp. 407-414, 2004.
17. Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," *Proceedings of the 14th International Conference on Machine Learning*, Nashville, TN, 1997, July, pp. 412-420.
18. E. Charniak, Statistical Language Learning, Cambridge, MA: MIT Press, 1993.
19. S. Lee and J. Seo, "Korean speech act analysis system using hidden markov model with decision trees," *International Journal of Computer Processing of Oriental Languages*, vol. 15, no. 3, pp. 231-243, 2002.
20. D. Surendran and G. A. Levow, "Dialogue act tagging with support vector machines and hidden markov models," *Proceedings of Interspeech/ICSLP*, Pittsburgh, PA, 2006, Sep.
21. H. Lee, H. Kim, and J. Seo, "Efficient domain action classification using neural networks," *Neural Information Processing. Lecture Notes in Computer Science Vol. 4223*, I. King, J. Wang, L. W. Chan, and D. Wang, Eds., Heidelberg, Germany: Springer Berlin, 2006, pp. 150-158.
22. D. Kim, H. Kim, and J. Seo, "A statistical prediction model of speakers' intentions in a goal-oriented dialogue," *Journal of Korean Institute of Information Scientists and Engineers: Software and Applications*, vol. 35, no. 9, pp. 554-561, Sep 2008.
23. J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," *Proceedings of the Eighteenth International Conference on Machine Learning*, Williamstown, MA, 2001, June 28-July 1, pp. 282-289.
24. J. C. Reynar and A. Ratnaparkhi, "A maximum entropy approach to identifying sentence boundaries," *Proceedings of the 5th Conference on Applied Natural Language Processing*, Washington, DC, 1997, March 31-April 3, pp. 16-19.
25. J. R. Quinlan, C4.5: Programs for Machine Learning, San Mateo, CA: Morgan Kaufmann Publishers, 1993.
26. T. Joachims, "SVMLight: Support Vector Machine Version: 6.02," http://svmlight.joachims.org/.
27. E. S. Ristad, Maximum entropy modeling toolkit, Technical Report, Department of Computer Science, Princeton University, 1996.

Harksoo Kim

Harksoo Kim is an assistant professor of Computer and Communications Engineering at Kangwon National University. He received his BA in Computer Science from Konkuk University in 1996. He received his MS and PhD in Computer Science from Sogang University in 1998 and 2003, respectively. He visited the CIIR at the University of Massachusetts in Amherst as a research fellow in 2004. In 2005, he worked as a senior researcher at the Electronics and Telecommunications Research Institute. His research interests include statistical methods for natural language processing, dialogue management, relation extraction, and question answering.



Choong-Nyoung Seon

Choong-Nyoung Seon is a PhD candidate in computer science at Sogang University, Seoul. He received his BS and MS degrees in computer science from Sogang University in 1999 and 2001, respectively. He has worked as a senior researcher for four years at Diquest Inc., a major information retrieval company in Korea. His research interests include natural language understanding, information extraction, and natural language generation.

Jungyun Seo

Jungyun Seo is a professor of computer science at Sogang University. He was educated at Sogang University, where he obtained a BS in Mathematics in 1981. He continued his studies in the Department of Computer Science at the University of Texas, Austin, receiving his MS and PhD in Computer Science in 1985 and 1990, respectively. He returned to Korea in 1991 to join the faculty of the Korea Advanced Institute of Science and Technology in Taejon, where he led the Natural Language Processing Laboratory in the Computer Science Department. In 1995, he moved to Sogang University in Seoul and became a full professor in March 2001. His research interests include multi-modal dialogues, statistical methods for natural language processing, machine translation, and information retrieval.