Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

A reliable multidomain model for speech act classification ☆

Sangwoo Kang [a], Harksoo Kim [b,*], Jungyun Seo [c]

[a] Department of Computer Science, Sogang University, South Korea
[b] Program of Computer and Communications Engineering, Kangwon National University, Chuncheon-si, Kangwon-do 200-701, South Korea
[c] Department of Computer Science and Interdisciplinary Program of Integrated Biotechnology, Sogang University, South Korea

ARTICLE INFO

Article history:
Received 29 September 2008
Received in revised form 7 July 2009
Available online 3 September 2009

Communicated by P. Franti

Keywords:
Speech act classification
Dialogue domain detection
Multidomain dialogue

ABSTRACT

In a multidomain dialogue, identifying speech acts is not easy because of the problem of interference between input features. To overcome this problem, we propose a two-step model for speech act classification. In the first step, the proposed model detects a dialogue domain associated with an input utterance. In the second step, the proposed model determines the speech act of the input utterance by using only statistical information about input features in the detected dialogue domain. In the experiment, the precision of the proposed model was higher than that of the baseline system without domain selection by 5.5%. On the basis of this experimental result, we conclude that reducing the interferences between input features by using a domain detection process is effective in improving the precision of speech act classification in multiple domains.

1. Introduction

A dialogue system is a software program that enables a user to interact with a system using natural language. An essential task of the dialogue system is to understand what the user says. For this, the system should identify speech acts that imply the user's intentions. Austin (1962) observed that there are three acts performed whenever something is said: a locutionary act, an illocutionary act, and a perlocutionary act. In this paper, a speech act means an illocutionary act (*i.e.*, a linguistic term used to describe an utterance that attempts to affect the addressee; such terms include commands, suggestions, inquiries, vows, and so on). Generally, a speech act represents a domain-independent general intention that is expressed in the form of an utterance. However, sometimes, domain-dependent words provide informative clues that help to identify a speech act. Therefore, many multidomain dialogue systems perform domain detection before speech act identification because users can switch dialogue domains according to their preferences. By using the results of domain detection, multidomain dialogue systems attempt to increase the accuracy of speech act identification.

Some initial approaches for speech act identification were based on knowledge such as recipes for plan inference and domain-specific knowledge (Carberry and Lambert, 1999; Lambert, 1993; Lee, 1996; Litman and Allen, 1987). Since these knowledge-based models depend on costly handcrafted knowledge, it is difficult to expand these models to more complex domains. Various machine learning models have been utilized to identify speech acts in order to overcome such problems (Choi et al., 1999; Kim et al., 2004; Lee and Seo, 2002; Samuel et al., 1999). Machine learning is a means to associate features of utterances with particular speech acts. Models based on machine learning have attracted attention since computers can automatically analyze large quantities of data and consider many different feature interactions. However, since previous machine learning models were focused on obtaining high performance in a specific domain, they did not deal with the problem of feature interference that frequently arises in a multidomain dialogue. If we train a speech act classification model by using mixed features without considering the differences between dialogue domains, the model will not perform satisfactorily because the mixed features generally include noisy and non-optimal information. In order to overcome this problem, we propose a two-step model for speech act classification in a multidomain dialogue. The proposed model first detects the dialogue domain associated with an input utterance. Then, the proposed model classifies the input utterance into a proper speech act category on the basis of statistical information about the detected domain. Some recent models have begun to deal with speech act identification problems in a multidomain dialogue (Lin et al., 2001; O'Neill et al., 2004). The previous models indicated that domain detection should be preceded by speech act identification. However, the previous models have a drawback in that domain-independent information can be missed because these models are simple keyword spotting

☆ This paper is an extended version of "Efficient Speech Act Analysis using Multiple Statistical Models and Domain Detection for Multiple Domains" in Proceedings of IEEE RO-MAN 2007.

* Corresponding author. Tel.: +82 33 250 6388; fax: +82 33 252 6390.
E-mail addresses: swkang@sogang.ac.kr (S. Kang), nlpdrkim@kangwon.ac.kr (H. Kim), seojy@sogang.ac.kr (J. Seo).

0167-8655/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2009.08.013

methods and use only the information that is specific to a detected domain without considering the accuracy of domain detection. In order to resolve this problem, the proposed model checks the confidence score of domain detection. If the confidence score is lower than a predefined threshold value, the proposed model attempts to identify a speech act by using statistical information collected from all domains.

2. Speech act classification in a multidomain dialogue

Given a dialogue $U_{1,n}$ that consists of n utterances, let $S_{1,n}$ and $D_{1,n}$ denote the speech acts and the dialogue domains of $U_{1,n}$, respectively. Then, an integrated model for simultaneously determining dialogue domains and speech acts can be formally defined as Eq. (1).

$$\begin{aligned}\mathrm{IM}(U_{1,n}) &= \underset{D_{1,n},S_{1,n}}{\arg\max} P(D_{1,n}, S_{1,n}|U_{1,n}) \\ &= \underset{D_{1,n},S_{1,n}}{\arg\max} P(D_{1,n}|U_{1,n})P(S_{1,n}|D_{1,n}, U_{1,n}) \qquad (1)\end{aligned}$$

In order to simplify computations and make performance tuning easy from a practical viewpoint, we divide Eq. (1) into two sequential equations, as shown in Eqs. (2) and (3). Many multidomain dialogue systems are developed by integrating single-domain systems in order to reduce the implementation effort. In these cases, domain detection is performed at the early stage of dialogue management because a dialogue manager has to load a dialogue model that is dependent on a specific dialogue domain. Therefore, the proposed model first detects a dialogue domain by using Eq. (2). Using the results of domain detection, it then determines a speech act, as shown in Eq. (3).

$$DM(U_{1,n}) = \underset{D_{1,n}}{\arg\max} P(D_{1,n}|U_{1,n}) \qquad (2)$$

$$SA(U_{1,n}) = \underset{S_{1,n}}{\arg\max} P(S_{1,n}|D_{1,n}, U_{1,n}) \qquad (3)$$

In this paper, we refer to Eqs. (2) and (3) as the domain detection model and the speech act classification model, respectively.

2.1. Domain detection model

As shown in Eq. (4), we simplify Eq. (2) by making the following assumption: a current domain is only dependent on a current utterance because users can switch domains at any time according to their preferences.

$$DM(U_{1,n}) = \underset{D_{1,n}}{\arg\max} \prod_{i=1}^{n} P(D_i|U_i) \qquad (4)$$

In Eq. (4), it is impossible to directly compute $P(D_i|U_i)$ because a speaker expresses identical content by using various surface forms of sentences according to his/her personal linguistic sense in a real dialogue. In order to overcome this problem, we approximate $P(D_i|U_i)$ by using the sentential feature set F_i, as shown in Eq. (5).

$$DM(U_{1,n}) = \underset{D_{1,n}}{\arg\max} \prod_{i=1}^{n} P(D_i|F_i) \qquad (5)$$

In Eq. (5), the sentential feature set consists of lexical features (content words annotated with POS's (parts-of-speech)) and POS features (POS bi-grams of all words in an utterance) (Kim et al., 2004; Kim et al., 2008). Generally, content words include nouns, verbs, adjectives, and adverbs, while functional words involve prepositions, conjunctions, and interjections. After extracting sentential features from utterances by using a conventional morphological analyzer, we remove non-informative features based on a well-known χ^2 statistic that measures the lack of independence between a feature and a category (in this paper, a dialogue domain or a speech act) (Kim et al., 2004). Then, we estimate Eq. (5) using CRFs (conditional random fields; a probabilistic undirected graph model for sequence labeling). The reason for using CRFs is that they have been applied to a number of natural language processing tasks with considerable success (Lafferty et al., 2001).

2.2. Speech act classification model

As shown in Eq. (6), we simplify Eq. (3) by making the following assumptions: a current speech act is only dependent on previous speech acts, and a current sentential feature set in a current dialogue domain contains informative clues for determining a current speech act.

$$SA(U_{1,n}) \approx \underset{S_{1,n}}{\arg\max} \prod_{i=1}^{n} P(S_i|D_i, F_i)P\big(S_i|S_{1,i-1}\big) \qquad (6)$$

In Eq. (6), since it is impractical to compute $P\big(S_i|S_{1,i-1}\big)$ by considering all preceding speech acts, we use a bi-gram model, as shown in Eq. (7).

$$SA(U_{1,n}) \approx \underset{S_{1,n}}{\arg\max} \prod_{i=1}^{n} P(S_i|D_i, F_i)P(S_i|S_{i-1}) \qquad (7)$$

In Eq. (7), the conditional probabilities on the right-hand side are effectively estimated by using CRFs in the same manner as the domain detection model, except that the speech act classification model has multiple statistical bases. If an application domain can be divided into n dialogue domains, we generate n statistical models for each dialogue domain, $n + 1$ statistical models using CRFs, and a statistical model for the entire application domain that contains the dialogue domains.

The speech act classification process consists of two steps. In the first step, the proposed model performs domain detection according to Equation (5). If the output score of the domain detection model is greater than a predefined threshold, the proposed model loads a statistical model of the detected domain and performs speech act classification according to Eq. (7). Otherwise, the proposed model ignores D_i in Eq. (7), loads a statistical model of the entire application domain, and performs speech act classification.

3. Experiments

3.1. Data sets and experimental settings

We collected three domains of a Korean dialogue corpus such as favorite foods (3009 utterances; 972 unique words), views on love (3092 utterances; 1874 unique words), and favorite music (1001 utterances; 655 unique words) from mobile chat rooms in which two users discuss each other's views on a specific topic by using the short message service of a commercial telecommunication company. Each utterance in the collected dialogues is manually annotated with speech acts. The speech acts were manually tagged by three undergraduate students who were familiar with dialogue analysis. Before manual tagging, we explained the meanings of speech acts to the students and showed them some samples that were annotated with correct speech acts. We spent approximately 2 h in training the students. Then, we assigned one student to each domain as a coder. In order to measure an agreement rate between coders, we calculated Fleiss' Kappa value (Fleiss, 1971) from 500 sampling utterances. A Fleiss' Kappa value of 0.75 indicates substantial agreement, although it is by no means universally accepted (Landis and Koch, 1977). The entire agreement rate for all coders was 60.4%. In order to evaluate the proposed model, we first divided the annotated dialogue corpus into a training corpus and a test corpus; the ratio of division was 4:1 for each dialogue domain. Second, we trained the proposed model by using the training corpus. Third, we manually disjointed the test corpus to prepare a list of discourse segments, where each discourse segment has a subtopic and comprises some dialogue turns that are pairs of a prompt from one speaker and a response received from the other speaker. Then, we randomly mixed the discourse segments without considering the dialogue domains. Finally, we performed 5-fold cross validation by using the training corpus and the domain-mixture corpus. The classification of speech acts is very subjective, and a universally agreed criterion does not exist. In this paper, we defined 40 types of speech acts. Fig. 1 shows the distribution of speech acts that occurred in the collected dialogue corpus.

For domain detection, we experimentally set the cut-off point of χ^2 statistic (the numbers of sentential features) as the top 1700 features of 2692 features. For speech act classification, we set the cut-off points as the top 700 of 1162 features in the food domain, the top 700 of 2005 features in the love domain, the top 400 of 828 features in the music domain, and the top 1700 of 2692 features in the entire application domain. We used the L-BFGS algorithm (Nocedal, 1980) to guess the internal arguments of CRFs and used Gaussian Prior to find a smoothing factor for the sparse data problem. We set Gaussian Prior to 10 and conducted the training 30 times. Then, we set the threshold value of domain detection to 0.7.

3.2. Experimental results

In order to evaluate the proposed model, we implemented a baseline model that determines speech acts by using sentential features of the entire application domain without domain detection. Table 1 shows the difference between the performance of the proposed model and that of the baseline model.

In Table 1, the average precision of domain detection was 77.4% for all utterances (87.6% for the utterances over the threshold value). As shown in Table 1, the precision of the proposed model was higher than that of the baseline model by 5.5%. We speculate that this increase is caused by the use of appropriate features. In the training corpus of the food domain, many food names such as "spaghetti" and "beef steaks" occurred, and utterances including the food names were mainly annotated with some speech acts like "response-what" and "response-which." However, we could not find these situations in the training corpus of the other domains. As a result, the food names were excluded from the lexical features of the baseline model because they yielded low χ^2 values. In addition, some words were used as indicators for different speech acts according to the dialogue domains. For example, "love" was highly associated with "statement" in the love domain, for example, "Love is a helping hand." However, it was associated with "response-what" in the food domain, for example, "I love spaghetti." After considering all these facts, we conclude that many informative and domain-specific lexical clues can be used to increase the precision of speech act classification in each domain. In the experiment, the average number of domain-presumed utterances (*i.e.*, utterances that obtain scores over the threshold value in the domain detection) was 1115 (approximately 81.6% of the test corpus). The average precision of speech act classification on the domain-presumed utterances was 94.9%. We found domain-specific lexical features in many domain-presumed utterances that are correctly classified by the proposed model but are incorrectly classified by the baseline model. Therefore, we believe that the precision of the models was affected by domain-specific features. The average precision (86.2%) of speech act classification of domain-ignored utterances (*i.e.*, utterances that obtain scores under the threshold value in the domain detection) was a little lower than the average precision (87.8%) of the baseline system. Further, we speculate that the lower precision was caused by the difference in test data. On the basis of these experimental results, we conclude that although the domain detection module does not perform well, reducing the interferences between input features by using a domain detection module helps to improve the precision of speech act classification in a multidomain dialogue.

Table 1
Difference in performances.

Model	Classification target	Average precision (%)
Baseline model	All utterances	87.8
Proposed model	Domain-ignored utterances	86.2
	Domain-presumed utterances	94.9
	All utterances	93.3

We analyzed the cases in which the proposed model failed to return correct results. The reasons for failure are as follows: first, we found some cases (1.86% of the test corpus) in which incorrect domain detection gave rise to speech act classification errors. In order to resolve this problem, we plan to study methods for improv-

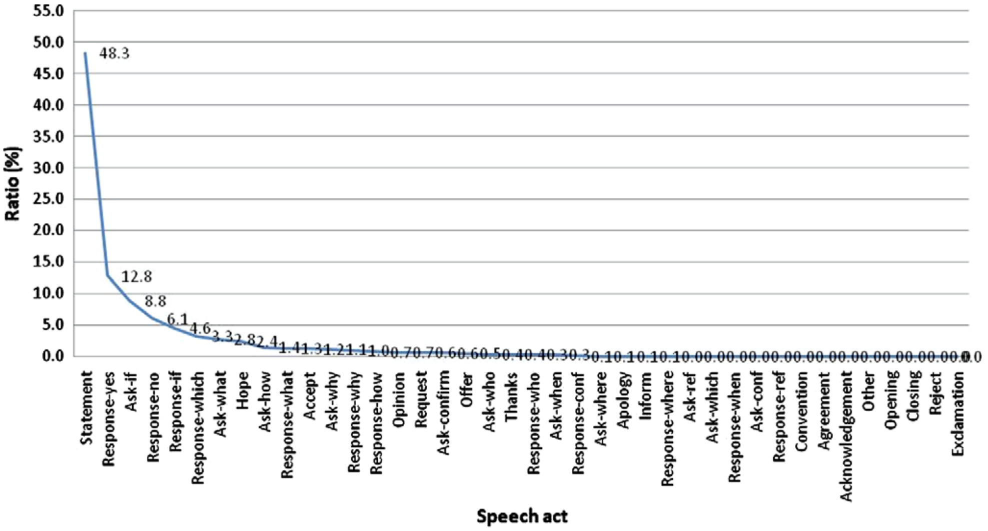

Fig. 1. Distribution of speech acts.

ing the precision of domain detection. Second, in some cases (4.80% of the test corpus), sparseness of the training corpus resulted in speech act classification errors. To overcome this issue, we shall study smoothing methods such as the backed-off model. Finally, some cases (0.04% of the test corpus) existed in which the ambiguity of the speech act categories led to speech act classification errors. Sometimes, an utterance was associated with several speech acts. To overcome this problem, we shall study methods for assigning an utterance to multiple speech acts.

4. Conclusion

We proposed a two-step model for speech act classification in a multidomain dialogue. In order to overcome the problem of interference between input features in the multidomain dialogue, the proposed model first detects dialogue domains of input utterances. Then, the proposed model determines the speech acts of the input utterances by using only statistical information about input features in the detected domain. In the experiment, the precision of the proposed model was higher than that of the baseline model without domain detection by 5.2%. On the basis of this experiment, we found that it is important to reliably extract input features according to each dialogue domain for speech act classification in a multidomain dialogue.

Acknowledgement

This work was supported by the Special Research Grant of Sogang University.

References

Austin, J.L., 1962. How to Do Things with Words. Oxford University Press, New York.

Carberry, S., Lambert, L., 1999. A process model for recognizing communicative acts and modeling negotiation subdialogues. Comput. Linguist. 25 (1), 1–53.

Choi, W., Cho, J., Seo, J., 1999. Analysis system of speech act and discourse structures using maximum entropy model. In: Proc. 27th Annual Meeting of the Association for Computational Linguistics, pp. 230–237.

Fleiss, J.L., 1971. Measuring nominal scale agreement among many raters. Psychol. Bull. 76 (5), 378–382.

Kim, K., Kim, H., Seo, J., 2004. A neural network model with feature selection for Korean speech act classification. Int. J. Neural Syst. 14 (6), 407–414.

Kim, M., Park, J., Kim, S., Rim, H., Lee, D., 2008. A comparative study on optimal feature identification and combination for Korean dialogue act classification. J. KISS: Softw. Appl. 35 (11), 681–691 (in Korean).

Lafferty, J., McCallum, A., Pereira, F., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. 18th Internat. Conf. on Machine Learning, pp. 282–289.

Lambert, L., 1993. Recognizing Complex Discourse Acts: A Tripartite Plan-based Model of Dialogue. Ph.D. Thesis, The University of Delaware.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

Lee, H., 1996. Analysis of Speech Act for Korean Dialogue Sentences. M.S. Thesis, Sogang University.

Litman, D., Allen, J., 1987. A plan recognition model for subdialoguesin conversations. Cognit. Sci. 11, 163–200.

Lee, S., Seo, J., 2002. A Korean speech act analysis system using hidden markov model with decision trees. Internat. J. Comput. Process. Orient. Lang. 15 (3), 231–243.

Lin, B., Wang, H., Lee, L., 2001. A distributed agent architecture for intelligent multi-domain spoken dialogue systems. IEICE Trans. Info. Syst. E84-D (9), 1217–1230.

Nocedal, J., 1980. Updating quasi-newton matrices with limited storage. Math. Comput. 35, 773–782.

O'Neill, L., Hanna, P., Liu, X., McTear, M., 2004. Cross domain dialogue modeling: An object-based approach. In: Proc. 8th Internat. Conf. on Spoken Language Processing, pp. 205–208.

Samuel, K., Carberry, S., Vijay-Shanker, K., 1999. Automatically selecting useful phrases for dialogue act tagging. In: Proc. 4th Conf. of the Pacific Association for Computational Linguistics.