PREDICTING AND TAGGING DIALOG-ACT USING MDP AND SVM

Keyan Zhou[1], Chengqing Zong[2], Hua Wu[3], Haifeng Wang[4]
[1,2]NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190
[3,4]Toshiba(China) Research and Development Center, Beijing, 100738
[1,2]{kyzhou, cqzong}@nlpr.ia.ac.cn; [3,4]{wuhua, wanghaifeng}@rdc.toshiba.com.cn

ABSTRACT

Dialog-act tagging is one of the hot topics in processing human-human conversation. In this paper, we introduce a novel model to predict and tag the dialog-act, in which Markov Decision Process (MDP) is utilized to predict the dialog-act sequence instead of using traditional dialog-act based *n*-gram, and Support Vector Machine (SVM) is employed to classify the dialog-act for each utterance. The predicting result of MDP and the classifying result of SVM are integrated as the final tagging. The experimental results have shown that our approach outperforms the traditional method.

Index Terms— MDP, Dialog-Act Modeling, SVM

1. INTRODUCTION

Dialog-act reflects the intention of a speaker, which is also a description of utterance with respect to syntactic, semantic, and pragmatic information. Dialog-act is widely used in speech recognition [3] and spoken dialog system [13]. The automatic recognition of dialog-act is one of the key problems in spoken language understanding.

In the recent years, new machine learning methods have been employed to tag the dialog-act, which have greatly improved the accuracy. However, these models classify dialog-act based on separate utterance, while little work has been done to incorporate the contextual information since dialog-act based *n*-gram and HMM of conversation were proposed in [6].

In this paper, we propose a novel approach to modeling dialog-act. In our approach, the contextual information is modeled by an MDP in order to predict dialog-act sequence instead of traditional methods. A popular tool of SVM is employed to classify the dialog-act for each utterance.

The remainder of the paper is organized as follows: Section 2 introduces the related work; Section 3 gives the details of our motivations and implementation; the experiments are shown in Section 4 and the concluding remarks are given in the final Section 5.

2. RELATED WORK

Generally, modeling the dialog-act consists of two processes: dialog-act tagging and discourse structuring.

For the dialog-act tagging, the previous work is usually based on the maximum entropy methods and decision tree classifiers [9] [1]. In the recently years, new techniques on classifier have been used successfully. For example, SVM is used in [10], and graphical model is used in [4]. However, most of them can not incorporate contextual information.

For discourse structuring, HMM and dialog-act based *n*-gram are used to do discourse structuring. Reference [9] mentioned that the sequence of dialogue acts should be constrained via a dialog-act based *n*-gram. [10] used a combination of SVM and HMM for dialog-act tagging, and obtained better result than those previously reported.

The graphical model is also a popular technique to do discourse structuring. [4] gave a conclusion that using the previous utterance does not help to predict the dialog-act of current utterance, which differs from [7]. [7] declared their graphical model significantly outperforms the HMM equivalent in the task of topic modeling, because HMM cannot combine different topics.

According to the existing research results, we have a consensus that the contextual information is very helpful in dialogue modeling. The contextual information for discourse structuring contains not only dialog-act sequence, but also topic information, speaker changes, as well as speaker identities and roles, which are hard to be described by using traditional models.

3. OUR MOTIVATIONS

3.1. Motivations

Dialog-act modeling, especially discourse structuring task, is an open question for its complex feature selection and integration; corpus collection and annotation are also expensive and time-consuming. Our aim is to construct a dialog model for dialog-act predicting task, which will be applied to dialog-act tagging task finally.

Let's see a dialog from another perspective: if the computer represents a participant in the dialog, knowledge of dialog states can be assumed, as in [11]. The task of dialog-act predicting turns to be a decision process.

MDP is a model for sequential decision process, which can easily integrate multiple features and make decision based on them. MDP is widely used in spoken dialog system for searching the optimal strategy while the system interacts with users, such as in [5] and [12]. However, in human-

human dialog-act predicting, very little previous work has been done with this model.

SVM is a widely used classifying technique, for which libsvm is a well-known tool [2]. We choose libsvm-2.84 as baseline for its convenience and utility.

Our framework consists of there models: dialog-act predicting based on MDP; dialog-act tagging based on SVM; the combination of MDP prediction and SVM classification.

3.2. Cast Dialog-Act Predicting as an MDP

Formally, an MDP is defined as a tuple $\{S, A, T, R\}$, which is formalized as follows:

- *State Descriptions(S)*

Dialog state is composed of two parts: speaker and dialog-acts. Speaker describes speaker's changing in a conversation, denoted as *sp_change* in Table 1. Dialog-acts record the last dialog-act of each speaker, which can be represented as {*DA_pre*, *DA-other*} or {*DA_pre*, *DA_self*} depending on *sp_change*, as shown in Table 1.

State	Values
sp_change	if speaker changes, sp_change=1; else, sp_change= 0
DA_pre	dialog-act of the previous utterance
DA_other	if sp-change == 0: dialog-act of the other speaker in the previous turn
DA_self	if sp-change == 1: dialog-act of the same speaker in the previous turn

Table 1. MDP State Features.

- *Action Set(A)*

There are totally 13 actions; each represents a dialog-act tag, as the prediction of the next utterance. See Appendix A.

- *Transition Probabilities(T)*

T is a matrix, in which $T_{ij}=P(S_j|S_i,A_i)$.

$P(S_j|S_i,A_i)$ denotes the probability of transition from S_i to S_j given an action A_i. Since the next state S_j simply depends on S_i and A_i, $P(S_j|S_i,A_i)$can be represented as:

$$P(S_j \mid S_i, A_i) \Leftrightarrow P(A_i \mid S_i, A_i) = P(A_i \mid S_i)$$

As state defined, T_{ij} can be further simplified as:

$$T_{ij} = \begin{cases} P(A_i \mid DA\_pre, DA\_other), & if\ sp\_change = 0 \\ P(A_i \mid DA\_pre, DA\_self), & if\ sp\_change = 1 \end{cases}$$

However, we don't know $P(A_i|DA\_pre,DA\_other)$ or $P(A_i|DA\_pre,DA\_self)$, but the posterior probability $P(DA\_pre,DA\_other|A_i)$ and $P(DA\_pre,DA\_self|A_i)$ can be calculated for each A_i. According to the Bayesian rule:

$$P(A_i \mid DA\_pre, DA\_other) = \frac{P(DA\_pre, DA\_other \mid A_i)P(A_i)}{P(DA\_pre, DA\_other)}$$

For each *DA_pre* and *DA_other* pair, the denominator $P(DA\_pre, DA\_other)$ is a fix value, which does not depend on A_i, and so it suffices to choose the A_i that maximizes the numerator. Thus, we can simplify T_{ij} as:

$$T_{ij} \propto \begin{cases} P(DA\_pre, DA\_other \mid A_i)P(A_i), & if\ sp\_change = 0 \\ P(DA\_pre, DA\_self \mid A_i)P(A_i), & if\ sp\_change = 1 \end{cases}$$

- *Reward(R)*

The reward function is a reflection of whether the prediction is right or not. We give an empirical reward or punishment to the *Transition Probabilities(T)* matrix. If the prediction is right, we assign a reward coefficient of ($\times 1.1$), or we assign a punishment coefficient of ($\times 0.9$), which are optimal values in a series of tests.

3.3. Dialog-Act Classifying with SVM

- Task description and model: Dialog-act classification is a multi-category classification task. We define 13 categories of dialog-acts as shown in Appendix A. The tool we use here is libsvm-2.84 [2].
- Features: we use three kinds of features: speakers, words, and punctuations. We focus on how to combine SVM and MDP in dialog-act tagging, rather than selecting effective features for SVM. Therefore, we do little discussion on feature selection.

3.4. Combination of Prediction and SVM Classification

There are two methods to combine the results of SVM classification and MDP Prediction:

- Using SVM classifying result to weigh MDP transition probabilities. We first get classifying results with probability estimation using libsvm. Then we assign SVM probability estimation to the MDP transition probability matrix to weigh each predicting result.
- Training MDP predicting result as a feature of SVM. Similarly, we get the MDP predicting result with probability, and put it into SVM feature spaces. We use this method in our experiment because of its convenience.

n-gram based dialog-act model is a typical traditional method for modeling dialog-act, which is discussed in detail in [9]. We experimented in predicting dialog-act with bi-gram based dialog-act model, and put the prediction into SVM feature spaces in the same way as in MDP.

4. EXPERIMENT AND ANALYSIS

4.1. Data and Labels

We collected Chinese human-human dialogues in text covering five domains. Hotel-Reservation is transcribed from conversational telephone speech and manually corrected. Others are collected from traveling booklets. The size of each corpus is described in Table 2.

We found there are a lot of similarities between our corpus and ICSI MRDA [8], so our dialog-act tag set is modified from MRDA labeling guide [3]. There are two levels of dialog-act tags: general tags which represent the basic form of an utterance (e.g., statement, question, etc.), and specific tags which are appended to the general tags. Specially, considering the ill-formedness in conversations, we add tag set called interruption, which contains 3 tags

Domain	Conversations	Utterances
Hotel-Reservation	171	6,094
At-the-Restaurant	45	653
Shopping	70	1,127
Taking-Taxi	33	441
At-the-bank	63	919
Total	382	9,234

Table 2. Capacity of corpus in each domain.

(abandoned, interrupted, and indecipherable). Each utterance contains one general tag or interruption tag.

The automatic tagging set contains the 10 general tags and 3 interruption tags, 13 tags in total (see at Appendix A).

4.2. Performance of SVM Classifier

We conducted two kinds of experiments. One is for domain adaptation, where the training data and test data were of different domains, and the other is for in-domain test, where these data sets were from the same domains. For the latter case, we also performed two experiments. One is five-fold cross validation experiment and the other is closed test, where the testing data is included in the training data. Results judged as precision are shown in Table 3, of which in diagonal line there are five-fold cross validation results and closed test results.

Compared with Table 2, "H" stands for "**H**otel Reservation"; "R" for "At-the-**R**estaurant"; "**S**" for "**S**hopping"; "T" for "Taking-the-**T**axi"; "B" for "At-the-**B**ank"; number of utterances in corpora are in parentheses. We conclude from Table 3 that

- Precision increases when training data enlarges. Column 2 and column 3 give the comparison of precision affected by the size of training data. When the training data enlarges from 4624 utterances to 6094 utterances, the precision of each testing set increases significantly.
- When the training and testing data are similar, results get better. The corpus of "T" has less utterance than others apparently; yet as training set, it gets the highest precision in testing set "S" except five-fold cross validation. Relatively, "S" is the most suitable training set to "T": the result is even better than that of five-fold cross validation. When we put the two corpora together, precision of five-fold cross validation rises to 70.11%.
- Precision of closed test is apparently much higher than that of five-fold cross validation. However, in "H", precision is lower than that we have expected. That's because "H" is spontaneous conversation, which is different from other domains. It is easy to see that there are close relationship between utterances when communicating. If contextual information is abandoned, it is hard to get satisfactory results.

4.3. Performance of MDP Predicting

Results of MDP predicting are shown in column 2 of Table 4. In the domain "H", we did three tests using different training sets: "H", 70 conversations (2517 utterances) of "H", "H" together with "S".

Results show that in the same domain, precision is proportional to the quantity of the training data. However, the precision suffers when the training set is added by data from other domains. In other tests, training with "H" gets better results in all domains for its remarkable data size.

Generally speaking, the performance of MDP predicting lives up to our expectations. Possible further improvement might be obtained if we add topic information to *State Descriptions(S)* and do Minimum Error Rate Training to optimize *Reward(R)*. Furthermore, we need to collect and annotate more data in order to do effective training.

4.4. Results and Analysis on Combination of Prediction and SVM Classification

We get the MDP predicting result with probability, and put it into SVM feature spaces. Results are shown in column 4 of Table 4. Experiments validate the improvement of the performance after combing MDP prediction to SVM: comparisons on SVM 5-fold validation are shown in Column 3 and column 4. In each domain, performance gets better when MDP prediction is added.

We also combined bi-gram based dialog-act model [9] with SVM in the same way as MDP. Results are shown in column 5.

Apparently, in all testing set, MDP model gets better performance than bi-gram based dialog-act model which sometimes hurts the performance of SVM. However, when trained in different domains, bi-gram based dialog-act model sometimes gets the same or a little higher scores. We think bi-gram based dialog-act model might be better in domain adaptation, which needs further verification in future work.

Although MDP works for dialog-act tagging tasks under most circumstances, this series of experiments show two shortcomings of MDP. First, MDP needs huge amount of training data; second, MDP is weak in domain adaptation. MDP might have more applications if it can be easily adapted to different domains.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we propose MDP for predicting dialog-act sequence comparing with traditional methods, and analyze results of MDP and SVM in details.

For SVM classification, we conclude that precision increases with larger training set or better consistency between the training and test set. For MDP predicting, the results show that in the same domain, precision is proportional to the quantity of training set and suffers when the training data are from unsimilar domains.

Test \ Train	H(4624)	H(6094)	R(653)	S(1127)	T(441)	B(919)
H(6094)	----	**78.86**/82.16	69.40	70.38	69.59	69.36
R(653)	63.71	64.47	**68.76**/89.28	65.39	63.25	63.71
S(1127)	61.40	61.76	60.43	**68.88**/88.91	62.56	61.22
T(441)	65.31	66.21	63.95	**68.93**	67.12/89.80	62.13
B(919)	66.16	66.92	60.72	64.42	62.50	**73.56**/92.27

Table 3. Precision (%) of SVM Classifier.

Domain \ Model		MDP Predi-ction	SVM 5 fold	SVM 5fold + MDP	SVM 5fold + bi-gram
Test	Train				
H (6094)	H	**61.80**	78.86	**79.03**	78.99
	H2517	52.12		78.90	78.88
	H and S	58.76		78.98	**79.03**
R (653)	H	**55.13**	68.76	69.07	*68.61*
	R	49.62		**69.22**	*68.30*
S (1127)	H	**52.66**	68.88	69.77	69.06
	S	40.70		**69.95**	69.06
T (441)	H	**53.74**	67.12	*66.89*	**67.12**
	T	43.54		**67.57**	67.12
B (919)	H	**52.45**	73.56	74.32	*73.45*
	B	46.57		**74.65**	73.56

Table 4. Combination of Prediction and SVM (%).

Results of combining SVM and MDP show that MDP outperforms traditional models. However, MDP has two weaknesses: first, MDP requires huge training data; second, MDP is weak in domain adaptation.

In future work, there are several key directions to improve our system. For example: data collection and labeling; applying our methods on other corpora, such as ICSI-MRDA; domain adaptation for MDP; improvement of MDP model; and the combination of MDP and SVM.

6. ACKNOWLEDGMENTS

The research work has been funded by the Natural Science Foundation of China under Grant No. 60575043, and the Hi-Tech Research and Development Program ("863" Program) of China under Grant No. 2006AA01Z194 as well.

7. REFERENCES

[1] Ang, J., Y. Liu, and E. Shriberg. Automatic Dialog-Act Segmentation and Classification in Muliparty Meetings. In *Proc. of 2005 IEEE ICASSP,* Philadelphia. 2005.

[2] Chang, C. and C. Lin. 2001. LIBSVM : a Library for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm

[3] Dhillon, R., S. Bhagat, H. Carvey, and E. Shriberg. Meeting Recorder Project: Dialog-act Labeling Guide. ICSI Technical Report TR-04-002, International Computer Science Institute. 2004.

[4] Ji, G. and J. Bilmes. Dialog-act Tagging Using Graphical Models. In *Proc. of 2005 IEEE ICASSP,* Philadelphia. 2005.

[5] Levin, E., R. Pieraccini, and W. Eckert. A Stochastic Model of Human-Machine Interaction for Learning Dialog Strategies. *IEEE Trans. on Speech and Audio Processing*, 8(1): 11-23. 2000.

[6] Nagata, M. and M. Tsuyoshi. First Steps toward Statistical Modeling of Dialogue to Predict the Dialog-Act Type of the Next Utterance. *Speech Communication*. 15: 193-203. 1994.

[7] Purver, M., K. P. Körding, T. L. Griffiths, and J. B. Tenenbaum. Unsuperivised Topic Modeling for Multi-Party Spoken Discourse. In *Proc. of COLING 2006,* Sydney. 2006.

[8] Shriberg, E., R. Dhillon, S. Bhagat, J. Ang, and H. Carvey. The ICSI Meeting Recorder Dialog-act(MRDA) Corpus. In *Special Interest Group on Discourse and Dialogue (SIGdial)*, Boston. 2004.

[9] Stolcke, A., K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Matin, C. Ess-Dykema, and M. Meteer. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26(3):339-373. 2000.

[10] Surendran, D., and G. Levow. Dialog-act Tagging with Support Vector Machines and Hidden Markov Models. In *Proc. of Interspeech*, Pittsburgh, PA. 2006.

[11] Taylor, P., S. King, S. Isard, and H. Wright. Intonation and Dialog Context as Constraints for Speech Recognition. *Language and Speech*, vol. 41: 489-508. 1998.

[12] Williams, J. and S. Young. Partially Observable Markov Decision Processes for Spoken Dialog Systems. *Computer Speech and Language*, 21(2):231-422. 2007.

[13] Wu, C., J. Yeh, and G. Yan. Dialog-act Modeling and Verification in Spoken Dialogue Systems. Advances in Chinese Spoken Language Processing. World Scientific, Chapter 14:321-339. 2007.

Appendix A.

Tags Compared with ICSI-MRDA and Percentage of Each

Tag	%age	Description	ICSI-MRDA
s	61.59	Statement	s
qy	20.09	Y/N Question	qy
qw	9.29	Wh-Question	qw
prt	3.69	Parenthesis	not marked
is	2.50	Imperative Sentence	not marked
%-	0.79	Interrupted	%-
qh	0.59	Rhetorical Question	qh
qo	0.46	Open-ended Question	qo
qr	0.40	Or Question	qr
%--	0.33	Abandoned	%--
es	0.13	Exclamatory Sentence	not marked
qrr	0.07	Or Clause After Y/N	qrr
%	0.07	Indecipherable	not marked