http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 8 (6): 923-928, 2009
ISSN 1812-5638

Improving Accuracy of Intention-Based Response Classification using Decision Tree

S.A. Ali, N. Sulaiman, A. Mustapha and N. Mustapha
Faculty of Computer Science and Information Technology, University Putra Malaysia,
43400 UPM Serdang, Selangor Darul Ehsan, Malaysia

Abstract: This study focused on improving the dialogue act classification to classify a user utterance into a response class using a decision tree approach. Decision tree classifier is tested on 64 mixed-initiative, transaction dialogue corpus in theater domain. The result from the comparative experiment show that decision tree able to achieve 81.95% recognition accuracy in classification better than the 73.9% obtained using Bayesian networks and 71.3% achieved by using Maximum likelihood estimation. This result showed that the performance of decision tree as classifier is well suited for these tasks.

Key words: Classification, decision tree, natural language generation, dialogue systems

INTRODUCTION

A dialog system is a computer system intended to converse with a human, with a coherent structure deal with interaction management issues such as turn-taking and topic management and with social aspects of communication like greeting and apologizing (Keizer and Bunt, 2007). In this kind of systems, one of the main concerns is the coherency of the response utterances. In contrast to speech recognition systems, where the goal is the correct transcription of the user utterances. This allows ignoring some words, focusing the attention on those which provide useful information for extracting the meaning of the utterance (Castro *et al*., 2004). This means, in response systems to generate a potential in the dialogue must be able to recognize this response, while maintaining equal semantic content.

Dialogue systems need to perform dialog act classification, in order to understand the role that an utterance plays in the dialog (e.g., a question for information or a request to perform an action) and to generate an appropriate next turn. In recent years, a variety of empirical techniques have been used to train the dialogue act classifier (Reithinger and Maier, 1995; Stolcke *et al*., 2000; Walker *et al*., 2001).

Yang *et al.* (2008) trained a decision tree classifier using prosodic features to show that the use of dialogue act tagging and prosodic information can help to improve the identification of action item descriptions and agreements. In addition to these features such as confidence score of action motivators and prosody, can be automatically extracted without costly human labeling.

Fodor (2007) presented the basis of a dialogue management communication mechanism that supports decision processes based on decision tree. Where learning the decision trees for conversations one can optimize the dialogue management and minimize the number of turn-takes steps in the dialogue. The decision tree represents the chronological ordering of the actions via the parent-child relationship and uses an object frame to represent the information state. The findings can be successfully applied in dialogue applications, such as contact center solutions.

Olguin and Cortés (2006) presented methodology promises a simple way to identify dialogue act types for the construction of dialogue managers. The methodology proposed CART-style decision trees on a corpus data where predictor data are utterance duration and sentence mood and the target data is the dialogue act type; first, sentence mood is predicted from INTSINT intonation taggings. The utility of predicting sentence mood was shown by comparing trees where tagged sentence mood, predicted sentence mood and no sentence mood at all were assessed. The resulting decision trees can be represented as if-then rule sets which can be programmed into a dialogue management system to identify the dialogue act type of an unknown utterance.

Komatani *et al.* (2005) proposed an abstract structure of a database search task and model it in two modes: specifying query conditions and requesting detailed

Corresponding Author: Sherif Abdulbari Ali, Faculty of Computer Science and Information Technology, University Putra Malaysia, 43400 UPM Serdang, Selangor Darul Ehsan, Malaysia
Tel: +60-3-8946-6585 Fax: +60-3-8946-6577

information. Then, define a set of very simple dialogue acts corresponding to the above dialogue model. Furthermore, they create a model to maintain query conditions as a tree structure, which can be used as a weight between attributes of query conditions. The constraints derived from these models are integrated by using a decision tree learning, so that the system can determine a dialogue act of the utterance and whether each content word should be accepted or rejected, even when it contains Automatic Speech Recognition (ASR) errors.

In classification-and-ranking, the decisions to choose from one response utterance over another require a considerable amount of domain knowledge. Hence, a knowledge-based approach as in deep generation is absolutely necessary. Deep generation determines the content of an utterance, or what to say, while the surface generation realizes the structure of the utterance, or determines how to say. Because deep generation requires a high degree of linguistic abstraction to produce fine-grained input specifications in order to drive the surface generators (Varges and Purver, 2006; Langkilde-Geary, 2002; Belz, 2007), its primary drawback is the classic problem of knowledge engineering bottleneck.

Overgeneration and ranking approaches to natural language generation have become increasingly popular (Paiva and Evans, 2005; Oh and Rudnicky, 2002). Overgeneration-and-ranking in dialogue processing performs mild overgeneration of candidate, followed by ranking to select the highest-ranked candidate as output (Varges and Purver, 2006). The main problem with this approach that it has to generate more candidates to form sentences. In addition to that, language models like n-gram have a built-in bias towards shorter strings is calculated as the likelihood of a string of words is the joint probability of the words. More precisely, the product of the probabilities of each word is given by n-1 preceding words (Belz, 2007). It is clear that this is not necessary for generation of dialogue utterances because all candidates must be treated equally, regardless of the length and the language rules.

In this study, we presented a response classification experiment based on user intentions using decision tree. The intention-based response generation systems require the task of classifying the response utterances into response classes. A response class contains all response utterances that are coherent to a particular input utterance. Classification-based NLG has been carried out for tasks in deep generation to guide the process of surface generation (Marciniak and Strube, 2004). However, as Stent (2002), former classification-based experiments do not take a full stochastic approach to response generation, but rather only in deep generation.

DIALOGUE CORPUS

The dialogue corpus SCHISMA (Schouwburg Informatie Systeem) is a collection of 64 text-based dialogues of a theater information and reservation system of tickets with the main objective of enabling users to make inquiries about theater performances scheduled and book a show of a wide range of options available. The corpus obtained through a series of Wizard of Oz experiments, built purposely for the acquisition of dialogue corpus for theater domain. The corpus contains 920 user utterances and 1127 server utterances in total. Schouwburg Infomatie System (SCHISMA) corpus is a mixed-initiative (Hulstijn and Van Hessen, 1998).

There are two types of interaction: inquiry and transaction. During inquiry the user has the initiative; the system answers the user's questions. When the user has indicated that he or she wants a reservation transaction the system takes initiative. The system will ask user series of questions like number of tickets to reserve, discount cards and others. User will answer the questions to complete the reservation details required by the system.

In transaction dialogue, before it reaches the stage of booking, the user and the system must cooperate to reach agreement on several issues such as the value of the ticket, the seating arrangement or the availability of discount. This model is more complex than the answer to the question systems because the system at any time, either party may request information from each other, especially for the user, it might come back out of any previous decisions and to start talking about a total opposite direction (Traum, 1997).

SCHISMA corpus is tagged using dialogue act annotation scheme based on Dialogue Act Markup in Several Layers (DAMSL) framework by Keizer and Op den Akker (2006). Table 1 lists the dialogue acts, represented as FLFs and BLFs in SCHISMA corpus. SCHISMA-DAMSL consists of five layers, each of which covers different aspect of communicative functions. This study concerned on two levels only, the forward-looking and backward-looking functions. Both levels indicate the communicative functions of an utterance. FLF tags indicate the type of speech act that the utterance is conveying, for example, assert, info-request and commit. BLF tags indicate how the particular utterance relates to the previous utterance and include answers (positive, negative or no-feedback) to questions, degree of understanding or disagreement.

Table 1: FLF and BLF for SCHISMA

FLF	User	System	BLF	User	System
Conventional	29	31	Signal_understanding	8	0
Commit	0	4	Signal_non_understanding	3	20
Offer	0	66	Positive_answer	162	399
Action_ directive	239	11	Negative_answer	30	42
Open_option	0	111	No_answer_feedback	3	63
Query_if	71	38	Correction_feedback	0	1
Query_ref	433	165	Accept, accept_part	70	54
Assert	123	694	Reject, reject_part	15	8
Exclamation	4	0	Hold	39	161
Explicit_ performative	2	0	Maybe	1	0
Other_ff	19	7	No_blf	589	379
	920	1127		920	1127

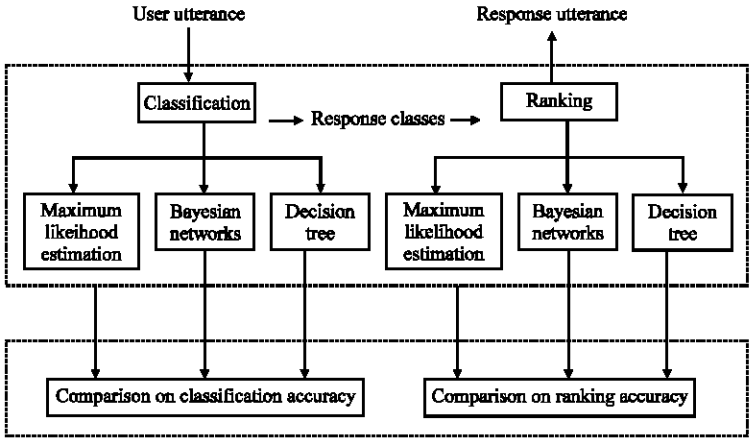

Fig. 1: The two-staged classification-and-ranking architecture

DECISION TREE FOR RESPONSE CLASSIFICATION

Decision tree learning is a method for approximating discrete-valued target functions, in which the learned function is represented by a decision tree. Learned trees can also be re-represented as sets of if-then rules to improve human readability. These learning methods are among the most popular algorithms and have been successfully applied to a broad range of tasks from learning to diagnose medical cases to learning to assess credit risk of loan applicants (Bar-Or *et al*., 2005). Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. Each node in the tree specifies a test of some attribute of the instance and each branch descending from that node corresponds to one of the possible values for this attribute (Mitchell, 1997).

Response classification is part of a two-staged classification-and-ranking architecture as shown in Fig. 1. This architecture proposed by Mustapha *et al*. (2008). The first component is a classifier that classifies user input utterances into response classes based on their contextual, pragmatic interpretations. The second component is a ranker that scores the candidate response utterances according to semantic content relevant to the input utterance.

One approach to classification would be to generate all possible decision trees that correctly classify the training set and to select the simplest of them. The number of such trees is finite but very large, so this approach would only be feasible for small classification tasks. ID3 decision tree was designed for the other end of the spectrum, where there are many attributes and the training set contains many objects, but where a reasonably good decision tree is required without much computation (Quinlan, 1986).

The basic structure of ID3 is iterative (Li and Aiken, 1998). A subset of the training set is chosen at random and a decision tree formed from it; this tree correctly classifies all instances in the subset. All other instances in the training set are then classified using the tree. If the tree gives the correct answer for all of these instances, it is correct for the entire training set and the process terminates. If not, a selection of the incorrectly classified instances is added to the subset and the process continues. This procedure will always produce a decision tree that correctly classifies each instance in the training

set, provided that a test can always be found that gives a nontrivial partition of any set of instances. For ID3, the choice of test is the selection of an attribute for the root of the tree. ID3 adopts a mutual-information criterion to choose that attribute to branch on that gains the most information. So, the inductive biases inherent in ID3 are preference biases that explicitly search for a simple hypothesis.

The main task of any classification task is to identify the set of classes that some observation belongs to, which is in this paper to identify a response class for each response utterances, such that P(response class|user utterance). The purpose of the response classification is to find the proper recognition for the accuracy of correct predictions of response class rc, given the user utterance U.

The user utterances are characterized by semantic and pragmatic features represented by nodes in the decision tree, at each node selecting the utterance properties that uniquely constitute the user utterance U that best classified. This process continues until the tree perfectly classified, or until all features have been used. We use rc to mean our estimate of the correct response class.

EXPERIMENTAL OPERATIONS

The experiments concerned on classification of user input utterances into response classes based on features extracted from user input utterances. The SCHISMA provides 920 instances of user utterance from 64 dialogues. The response class for each user utterance is manually tagged according to topic of the response utterances.

Tagging the response class adapts to patterns of input and response utterance per turn throughout the course of conversation to maintain the coherency in a sequence of two utterances. There were 15 response classes using the same naming conventions as topic in user utterance. Table 2 shows the statistics for the response classes.

The 10-fold cross validation is performed to split the data into ten approximately equal partitions, each being used in turn for testing while the remainder of data is used for training.

Table 3 shows the semantic and pragmatic feature used in the classification experiment. The speech acts FLF and grounding acts BLF from user utterances readily available from the DAMSL-annotated SCHISMA corpus.

We extend our experiment by testing another dialogue corpus in order to validate the firmness of the decision tree classifier. Corresponding to response classification experiment in SCHIMA corpus, we investigate the accuracy rate for response classification task with cross-domain experimentation using MONROE corpus as the secondary source of our validation experiment.

The MONROE corpus is a mixed-initiative interaction and collaborative problem-solving task in disaster scenario set in Monroe County, New York. Emergencies include car accidents, natural disasters such as flooding and snow storms, request for medical assistance, or civil disorders. Given a particular emergency task, the dialogue participants are expected to coordinate help for the task. The objects to coordinate for in this domain include people, roads, vehicles, crews and equipment (Stent, 2002).

Table 2: Statistics for response classes

Global topic	Response class	Frequency	Percentage
Performance	Title	104	11.3
	Genre	28	3.0
	Artist	42	4.6
	Time	32	3.5
	Date	90	9.8
	Review	56	6.1
	Person	30	3.3
Reservation	Reserve	150	16.3
	Ticket	81	8.8
	Cost	53	5.8
	Avail	14	1.5
	Reduce	73	7.9
	Seat	94	10.2
	Theater	12	1.3
Other	Other	61	6.6

Table 3: Features used as nodes in decision tree

Node name	Type	Values	Descriptions
Context	Scalar	{Performance, reservation}	Global topic of user utterance
Topic	Scalar	{Title, genre, artist, time, date, review, person, reserve, ticket, cost, avail, reduc, seat, theater, other}	Topic of conversation in user utterance
Action	Scalar	{Assert, question, command, other}	Classification of user utterance based on purpose i.e., declarative, interrogative or imperative
Control	Scalar	{Client, system}	Control holder at the point of user utterance
Role	Scalar	{Initiator, responder}	Role of the user
Turn	Scalar	{Release, take, keep}	Turn-taking act for user utterance
Negotiation	Scalar	{Open, inform, propose, confirm, close}	Negotiation act for user utterance
FLF	Scalar	Refer table 1	Speech act for user utterance
BLF	Scalar	Refer table 1	Grounding act for user utterance

Table 4: Response classification accuracy comparison

Corpus	Semantic features	Pragmatic features	Maximum likelihood estimation accuracy (%)	Bayesian networks accuracy (%)	Decision tree accuracy (%)
SCHISMA	Context, Topic	FLF, BLF, Action, Control, Turn, Role, Negotiation	71.3	73.9	81.9
MONROE	Context, Topic	FLF, BLF, Action, Control, Turn, Role, Negotiation	50.8	64.8	75.2

Table 5: Performance evaluation of the three classifiers

	Maximum likelihood estimation			Bayesian networks			Decision tree		
Corpus	Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
SCHISMA	0.711	0.729	0.719	0.823	0.805	0.813	0.887	0.856	0.871
MONROE	0.30	0.274	0.286	0.559	0.542	0.55	0.814	0.723	0.765

Same with response classification experiment in SCHIMA corpus, we investigate the accuracy rate for response classification task. The baseline accuracy for response classification in MONROE corpus is 64.8% achieved with Bayesian networks and 50.8% achieved with Maximum Likelihood. Table 4 shows the results of classification experiment.

DISCUSSION

Table 4 relates the results for response classification experiment from our approach using decision tree with the previous findings which are the Bayesian networks as well as Maximum Likelihood (Mustapha *et al*., 2008) using two sets of corpus, SCHISMA and MONROE. We achieved in accuracy result of maximum 81.95% better than the 73.9% obtained using Bayesian networks tested on the SCHISMA dialogue corpus and 71.3% achieved by using Maximum Likelihood estimation. Regarding to MONROE dialogue corpus, decision tree achieved 75.2% accuracy better than the two baseline approaches.

Table 5 shows the performance evaluations of the three classifiers, where the decision tree classifier performs better and achieves higher precision score compare to Bayesian network and Maximum Likelihood approach. The result shows empirical evidence that our approach using decision tree achieved the aim of the study on improving the dialogue act classification to classify a user utterance into response class. The decision tree correctly classified the user utterance with higher per cent recognition accuracy than both baseline approaches.

The result differs from previous study is due to the attribute selection measure where decision tree uses the information gain principles to take into account the discriminative power of each attribute over classes, in order to choose the best attribute one as the root then grow downward. This process ensures that the decision tree correctly classified the attributes. On the contrary, the previous work such as Bayesian networks, uses hill climbing search has no mechanism to alter the network structure to remove the arc at later stage, resulting in large number of conditional probabilities to be considered and often practically difficult to find significant features that optimize the classification performance in feature selection tasks. And another earlier study, Maximum Likelihood relies on rigid frequency counts, which often result in incorrect classification.

CONCLUSION

This study focused on classification of response utterances into response classes using decision tree. The experiment showed that the decision tree as classifier is well suited for classification task, where the classifier performances achieved the best accuracy of 81.95%. However, in order to improve the performance of decision tree classifier still further, it is clear that we need to search and analyze the 18.05% of user utterance that incorrectly classified.

REFERENCES

Bar-Or, A., D. Keren, A. Schuster and R. Wolff, 2005. Hierarchical decision tree induction in distributed genomic databases. IEEE Trans. Knowl. Data Eng., 17: 1138-1151.

Belz, A., 2007. Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. Natural Language Eng., 1: 1-26.

Castro, M., D. Vilar, P. Aibar and E. Sanchis, 2004. Dialogue Act Classification in a Spoken Dialogue System. Vol. 3040, Springer-Verlag, Berlin, Heidelberg, ISBN: 978-3-540-22218-7, pp: 260-270.

Fodor, P., 2007. Dialog management for decision processes. Proceedings of the 3rd Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, 2007, Poznan, Poland, pp: 1-4.

Hulstijn, J. and A. Van-Hessen, 1998. Utterance generation for transaction dialogues. Proceedings of the 5th International Conference of Spoken Language Processing, Nov. 30-Dec. 4, Sydney, Australia, pp: 1-4.

Keizer, S. and H. Bunt, 2007. Evaluating combinations of dialogue acts for generation. Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, 2007, Antwerp, Belgium, pp: 158-165.

Keizer, S. and R. Op-den-Akker, 2007. Dialogue act recognition under uncertainty using Bayesian networks. Natural Language Eng., 1: 1-30.

Komatani, K., N. Kanda, T. Ogata and H.G. Okuno, 2005. Contextual constraints based on dialogue models in database search task for spoken dialogue systems. Proceedings of the 9th European Conference on Speech Communication and Technology, Sept. 4-8, Lisbon, Portugal, pp: 877-880.

Langkilde-Geary, I., 2002. An empirical verification of coverage and correctness for a general-purpose sentence generator. Proceedings of the International Natural Language Generation Conference, 2002, IEEE Xplore, pp: 17-24.

Li, W. and M. Aiken, 1998. Inductive learning from preclassified training examples: An empirical study. IEEE Trans. Syst. Man Cybernet. Part C, 28: 288-294.

Marciniak, T. and M. Strube,, 2004. Classification-based generation using TAG. Proceedings of International Conference of National Language Generation, 2004, Brockenhurst, UK., pp: 100-109.

Mitchell, T., 1997. Machine Learning, Computer Science Series. McGraw Hill, New York, ISBN: 0070428077.

Mustapha, A., S. Nasir, R. Mahmod and H. Selamat, 2008. Classification-and-ranking architecture for response generation based on intentions. Int. J. Comput. Sci. Network Secur., 8: 253-257.

Oh, A.H. and A.I. Rudnicky, 2000. Stochastic natural language generation for spoken dialogue systems. Comput. Speech Language, 16: 387-407.

Olguin, S.R.C. and L.A.P. Cortés, 2006. Predicting Dialogue Acts from Prosodic Information. In: Computational Linguistics and Intelligent Text Processing, Gelbukh, A. (Ed.). LNCS., 3878, Springer-Verlag, Berlin, Heidelberg, ISBN: 8-3-540-32205-4, pp: 355-365.

Paiva, D.S. and R. Evans, 2005. Evans empirically-based control of natural language generation. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Jun. 25-30, Ann Arbor, Michigan, pp: 58-65.

Quinlan, J.R., 1986. Induction of decision trees. Machine Learn., 1: 81-106.

Reithinger, N. and E. Maier, 1995. Utilizing statistical dialogue act processing in Verbmobil. Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, Jun. 26-30, Morristown, NJ., USA., pp: 116-121.

Stent, A.J., 2002. A conversation acts model for generating spoken dialogue contributions. Comput. Speech Language, 16: 313-352.

Stolcke, A., K. Ries, N. Coccaro, E. Shriberg and R. Bates *et al*., 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics, 26: 339-373.

Traum, D., 1997. Views on mixed-initiative interaction. Proceedings of the AAAI97 Spring Symposium on Mixed-Initiative Interaction, Mar. 24-26, Stanford, CA., pp: 169-171.

Varges, S. and M. Purver, 2006. Robust language analysis and generation for spoken dialogue systems. Proceedings of the ECAI Workshop on Development and Evaluation of Robust Spoken Dialogue Systems for Real Applications, 2006, Italy, pp: 1-4.

Walker, M.A., R. Passonneau and J.E. Boland, 2001. Qualitative and quantitative evaluation of DARPA communicator dialogue systems. Proceedings of the 39th Annual Meeting of the on Association for Computational Linguistics. Jul. 06-11, Morristown, NJ., USA., pp: 515-522.

Yang, F., G. Tur and E. Shriberg, 2008. Exploiting dialogue act tagging and prosodic information for action item identification. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Mar. 31-Apr. 4, Las Vegas, NV., pp: 4941-4944.