# Dynamic Bayesian Networks in Classification-and-Ranking Architecture of Response Generation

Aida Mustapha, Md. Nasir Sulaiman, Ramlan Mahmod, Mohd. Hasan Selamat
Department of Computer Science, Faculty of Computer Science and Information Technology,
University Putra Malaysia, 43400 UPM Serdang, Selangor Darul Ehsan, Malaysia

**Abstract: Problem statement:** The first component in classification-and-ranking architecture is a Bayesian classifier that classifies user utterances into response classes based on their semantic and pragmatic interpretations. Bayesian networks are sufficient if data is limited to single user input utterance. However, if the classifier is able to collate features from a sequence of previous n-1 user utterances, the additional information may or may not improve the accuracy rate in response classification. **Approach:** This article investigates the use of dynamic Bayesian networks to include time-series information in the form of extended features from preceding utterances. The experiment was conducted on SCHISMA corpus, which is a mixed-initiative, transaction dialogue in theater reservation. **Results:** The results show that classification accuracy is improved, but rather insignificantly. The accuracy rate tends to deteriorate as time-span of dialogue is increased. **Conclusion:** Although every response utterance reflects form and behavior that are expected by the preceding utterance, influence of meaning and intentions diminishes throughout time as the conversation stretches to longer duration.

**Key words:** Bayesian networks, Conditional Probability Distributions (CPDs), dynamic Bayesian networks, Probabilistic Network Libraries (PNL), classification-and-ranking, Directed-Acyclic Graph (DAG), natural language generation, Natural Language Processing (NLP)

## INTRODUCTION

Response generation is a process of natural language generation in dialogue systems, which is responsible for providing dialogue responses as part of an interactive human-machine conversation. In human-human conversation, dialogue is mutually structured and timely negotiated between dialogue participants. Speakers take turns when they interact, they interrupt each other but their speeches seldom overlap. Each speaker is affected by what the other speaker has said and what each speaker says; affect what the next speaker will say. Similarly, human-machine conversation through dialogue systems must exhibit comparable qualities. But for dialogue systems to recognize turns, consider interrupts and maintain coherence, response generation must rely on pragmatic interpretation, apart from semantic understanding of user input utterances.

Classification-and-ranking generation is an alternative to grammar-based or statistical-based generation. This type of generation assumes that each user utterance represented in some context has its counterpart response in the dialogue corpus, hence promoting open-domain quality (Mustapha *et al.*, 2010). Classification-and-ranking architecture consists of two components: a Bayesian classifier to classify user utterances into response classes based on intentions of user input utterance and an Entropic ranker that scores the candidate response utterances according to semantics relevant to the user utterance (Mustapha *et al.*, 2008). The generation architecture is shown in Fig. 1.

Nonetheless, processing and generating natural languages requires understanding the interaction of complex knowledge sources, disguised in many forms. Much of what we understand about language is known with various degrees of certainty due to oversimplifying assumptions on many independencies and dependencies between context and meaning of language. Bayesian Networks (BN) is a natural choice of approach in Natural Language Processing (NLP) because it offers a formal treatment to uncertainties and independencies.

**Corresponding Author:** Aida Mustapha, Department of Computer Science,
Faculty of Computer Science and Information Technology, University Putra Malaysia,
43400 UPM Serdang, Selangor Darul Ehsan, Malaysia  Tel: +60-3-8947-1714, Fax: +60-3-8946-6577

Fig. 1: Classification-and-ranking generation

By using BNs, we are able to introduce the relationship of knowledge in two ways; within the structure of the network as well as the probability distributions of the network parameters. We can then reason under the uncertainty via the joint distributions.

In Natural Language Processing (NLP), Bayesian networks and dynamic Bayesian networks have long history of success in various tasks such as in dialogue act recognition (Keizer, 2003; Keizer and Akker, 2007; Ali *et al.*, 2007), word sense disambiguation (Chao and Dyer, 2000), speech recognition (Zweig and Russell, 1999), dialogue modeling (Pulman, 1996; Lemon *et al.*, 2002), human sentence processing (Narayanan and Jurafsky, 2002), question answering (Ramakrishnan *et al.*, 2003) and information retrieval (Ribeiro-Neto *et al.*, 2000), Representation, Inference and Learning (Murphy, 2002).

Bayesian networks have also gained significant attention in other domain such as in medical (Elsayad, 2010; Saat *et al.*, 2010) or intrusion detection (Khor *e*t *al.*, 2009; Mehdi *et al.*, 2007). Nonetheless, a BN is useful when the parameters in the domain are static, for example when each feature has a single and fixed value. However, in dialogue systems, utterances are produced in turns by two speakers, therefore data arrives sequentially turn after turn. Instead of analyzing one utterance at every turn, we can apply BN theory on the present and previous utterances from a sequence of turns. This leads to the proposed Dynamic Bayesian Networks (DBN) in response classification.

**Dynamic Bayesian networks:** Dynamic Bayesian Networks (DBNs) are directed graphical models of stochastic processes. A DBN is a specific type of BNs and consists of time-slices, whereby each time-slice or time-step contains its own network and variables. Given the time-variant property, however, a DBN must maintain the same network structure at each time-slice and the cross arcs are extended only between two consecutive time slices. Albeit the name of DBNs, the network structure really does not change. The term "dynamic" is to refer to time-dependent or sequence-dependent modeling and has nothing to do with the structure.

In a DBN, each position in the sequence-slice is characterized by n random variables. Within each slice, the random variables are represented by an ordinary BN, duplicated along the sequence positions. The sequential dependencies are represented by a set of arcs that connect the nodes across the consecutive sequence in the network. The topology of a DBN is a repeating structure and the Conditional Probability Distributions (CPDs) within each structure also do not change in each sequence-slice.

A DBN is defined as the pair $(B_1, B_\rightarrow)$ where $B_1$ is a BN that represent the prior state or initial state distribution of the state variables $(Z_1)$ (Ribeiro-Neto *et al.*, 2000). $Z_t = (U_t, X_t, Y_t)$, where $U_t$, $X_t$ and $Y_t$ represents the input, hidden and output variables of DBN. In the simplest problem where there are only two consecutive BNs, $B_\rightarrow$ is a two-slice temporal BN (2TBN) with transition as shown in Eq. 1:

$$p(Z_t|Z_{t-1})\prod_{i=1}^{N}p(Z_t^i|parents(Z_t^i)) \qquad (1)$$

where $Z_t^i$ is the *i*-th node at time t. Again, Z may represent either $U_t$, $X_t$ or $Y_t$. In turn, Parents $(Z_t^i)$ are the parents of $Z_t^i$, whether in the same or previous slice.

Since the structure repeats across the sequence of process, the parameters for slices t = 2, 3 and so forth remain the same. Note that nodes in the first slice of a DBN do not have parameters associated with them. Therefore, the parameters of the model can be fully described by only using the first two slices. JPD for the sequence of length T is obtained by "unrolling" the DBN as shown in Eq. 2:

$$p(Z_{1:T}) = \prod_{t=1}^{N}\prod_{i=1}^{N}p(Z_t^i|parents(Z_t^i)) \qquad (2)$$

In modeling a dialogue corpus, a sequence of user utterance is termed as sequence-slice in DBN, where t represents an utterance at time t. Hence, t = 0 represents the current utterance, t = 1 represents the previous one utterance, t = 2 represents the previous two utterances and t = 3 represents the previous three utterances.

## MATERIALS AND METHODS

Human being performs the task of classification in a variety of activity ranging from cognitive to behavioral tasks. We make decisions based on information available on hand and faithfully rely to such decisions yet in new but analogous situation. Once we settle with a particular situation, we will weigh and rank the options to make the best decision given the opportunities and constraints in that situation. In a

classification experiment, as in any classification task, the main task is to determine which of a set of classes some observation belongs to. In response classification, the objective is to identify a response class for each response utterances, maximizing $P$ (response class | user utterance). The list of response classes is shown in Table 1.

The experiment is to assess the classification accuracy of correct predictions of response class $rc$, given the user utterance $U$. The probability equation to find the best response class is given by Eq. 3:

$$\hat{rc} = \arg\max_{rc \in R} P(U|rc)P(rc) \tag{3}$$

Dynamic Bayesian Networks (DBN) model the sequence of user utterances in time-slices t, t-1, t-2 and so forth. This is possible through an extended set of semantic and pragmatic features that are extracted from a sequence of previous $n$-1 user utterances. In this experiment, we limit t to three previous utterances.

**Dialogue corpus:** The dialogue utterances under study are sourced from SCHISMA corpus. SCHISMA (SCHouwburg Informatie Systeem) is a Theater Information and Ticket Reservation system (Hoeven *et al.*, 1996). The dialogue corpus is a collection of 64 text-based, human-machine dialogues obtained through a series of Wizard of Oz experiments. It contains 920 user utterances and 1,127 server utterances in total. In total, there are 2,047 individual utterances in 1,723 turns.

SCHISMA is a mixed-initiative, transaction dialogue, in which there are two types of interaction: inquiry and transaction (Hulstijn, 2000). In transaction dialogue, both user and system must collaborate to achieve agreement on several issues like ticket price or discount availability before reaching the point of reservation. This model is more complex than question-answering systems because at any point, both parties may request information from each other and the user in particular, may retract any previous decisions and take the conversation in a totally different direction. The dialogue excerpts in Fig. 2 illustrate the complexities in the structure of mixed-initiative, transaction dialogues.

The dialogue commences with user browsing for information on theater performances. However, upon asking the ticket price as shown in Line 7, the system replies with a question to clarify on the presence of reduction card, which will give different ticket price altogether.

**Features:** Each utterance is analyzed from the perspective of speech actions, which is fully

| | | |
|---|---|---|
| U: | What will be on in the theatre next week (19 March): | [1] |
| S: | There is no show on that date | [2] |
| U: | And on 18 March? | [3] |
| S: | In the period 18 March 1994 until 20 March 1994 you can go to deelder denkit and indonseian tales | [4] |
| U: | At what time does deelder start? | [5] |
| S: | The show starts at 20:00 | [6] |
| U: | How much dose it cost | [7] |
| U: | and are there still places? | [8] |
| S: | Do you have a reduction card? | [9] |
| U: | No | [10] |
| S: | The price for the show deelder denkt is f26,00 | [11] |
| S: | And there are still 82 places free | [12] |

Fig. 2: Excerpt of SCHIMA corpus

Table 1: Statistics of response classes

| Response Class | Frequency | (%) |
|---|---|---|
| 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 |
| Reserve | 150 | 16.3 |
| Ticket | 81 | 8.8 |
| Cost | 53 | 5.8 |
| Avail | 14 | 1.5 |
| Reduc | 73 | 7.9 |
| Seat | 94 | 10.2 |
| Theater | 12 | 1.3 |
| Other | 61 | 6.6 |

Table 2: Semantic features used as nodes in DBN

| Node Name | Values | Descriptions |
|---|---|---|
| Context | {performance, reservation} | Global topic of user utterance |
| Topic | {title, genre, artist, time, date, review, person, reserve, ticket, cost, avail, reduc, seat, theater, other} | Topic in user utterance |

characterized by its (1) intentions and (2) semantic content in the form of input frame. These observed features are of utterance properties that uniquely constitute the user utterance, during a particular turn of a conversation. The SCHISMA corpus is readily tagged with DAMSL annotation scheme by Keizer and Akker (2007). In addition, there are two types of features extracted out from the user input utterances, which are semantic features and pragmatic features (Mustapha *et al.*, 2009) as shown in Table 2 and 3, respectively.

**DBN classification:** Training the SCHISMA corpus under Bayesian Networks (BN) and Dynamic Bayesian Networks (DBN) is carried out by using Probabilistic Network Libraries (PNL) (Intel, 2004). PNL support for dynamic Bayesian networks is the main reason to choose PNL over other statistical toolkit like WEKA (Ian and Frank, 2005). Structural learning in PNL is carried out using Hill-Climbing algorithm. This algorithm basically
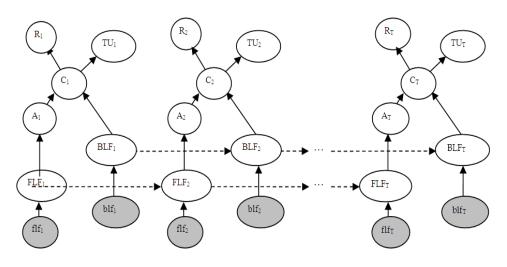
Fig. 3: An "unrolled" DBN

Table 3: Pragmatic features used as nodes in DBN

| Node Name | Values | Descriptions |
|---|---|---|
| Action | {assert, question, command, other} | Types of user utterance |
| Control | {client, system} | Control holder at user utterance |
| Role | {initiator, responder} | Role of the user |
| Turn | {release, take, keep} | Turn-taking act for user utterance |
| Negotiation | {open, inform, propose, confirm, close} | Negotiation act for user utterance |
| FLF | {conventional, commit, offer, action_directive, open_ option, query_if, query _ref, assert, exclamation, explicit_performance, other_ff} | Speech act for user utterance |
| BLF | {signal_understanding, signal_non_understanding, positive_answer, negative_answer, no_answer_feedback, accept, accept_part, reject, reject_part, hold, maybe, no_blf} | Grounding act for user utterance |

searches the space of Directed-Acyclic Graph (DAG) and builds the best arch to match training set based on the scoring function. Figure 3 illustrates one possible structure for an "unrolled" DBN produced by PNL. Figure 3 illustrates how the DBN is "unrolled" across the sequence of utterances, from $t = 0$ until $t = 3$. The dotted lines are tracking each particular feature variable. In testing accuracy of the DBN, a 10-fold cross validation was applied to split the SCHISMA corpus into ten approximately equal partitions training and testing set, each being used in turn for testing while the remainder combined for training.

## RESULTS

The goal of the experiment is to investigate the impact of features extracted from previous n user input utterances, if the semantic or pragmatic representation from the preceding utterances has any influence over the accuracy rate in classification of response utterance. The results are compared with Bayesian networks classification by (Mustapha *et al*., 2009). Table 4 shows comparison of accuracy percentages for response classification task using both semantic features and pragmatic features from previous user utterances.

## DISCUSSION

From Table 4, BN classification yields 73.9% accuracy percentage. As DBN classification is performed, results show that accuracy is improved, but rather insignificantly, either through time-series features in semantic content (semantic–*n*) or pragmatic (pragmatic-*n*). Even though previous topics contributed to increase in accuracy rate immediately with $t = 1$ and previous intentions only contributed to the increase in accuracy rate when $t = 2$, at the end the accuracy rate deteriorated as the time-span of dialogue increased.

This observation shows that intentions and semantic content in previous utterances do not gives enough impact to uniquely characterize input utterances. Figure 4 illustrates the changes in classification accuracy rates for the experiment. The *x*-axis shows time-series factor, where $t = 0$ represents the current utterance, $t = 1$ represents the previous one utterance, $t = 2$ represents the previous two utterance and $t = 3$ represents the previous three utterance. The *y*-axis shows classification accuracy in percentages.

Table 4: Comparison of response classification accuracy

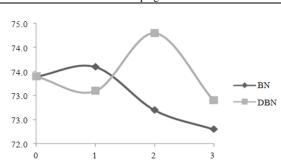| | | Accuracy (%) | |
| --- | --- | --- | --- |
| Semantic features | Pragmatic features | BN | DBN |
| Table 1 | Table 2 | 73.9 | - |
| BN and semantic-1 | - | - | 74.1 |
| BN and semantic-2 | - | - | 73.2 |
| BN and semantic-3 | - | - | 72.8 |
| - | BN and pragmatic-1 | - | 73.6 |
| - | BN and pragmatic-2 | - | 74.8 |
| - | BN and pragmatic-3 | - | 73.4 |



Fig. 4: Change of accuracy rates

## CONCLUSION

The underlying philosophy of classification-and-ranking architecture in natural language generation is for the response generator to directly learn response utterances from the domain corpus. The classification experiment was designed to classify response utterances into response classes in order to delimit the searching space for ranking the utterances. Through ranking, the response with highest probability will be returned as the final response to user input utterance.

The results for time-series experiments using dynamic Bayesian techniques are consistent with findings of dialogue act recognition in utterances (Ali *et al*., 2006), whereby consideration of previous n utterances in a dialogue does not necessarily affect the classification accuracy. While the first two previous utterances may increase the accuracy rate, the accuracy will deteriorate as the time-span of dialogue increased.

Although every response utterance reflects form and behavior that are expected by the preceding utterance, influence of meaning and intentions diminishes throughout time as the conversation stretches to longer duration.

## REFERENCES

Ali, A., R. Mahmod, F. Ahmad and M.N. Sulaiman, 2006. Bayesian networks model for dialogue act recognition based on intra-utterance features and inter-utterances context. Suranaree J. Sci. Technol., 13: 285-297. http://sutlib2.sut.ac.th/Sutjournal/Files/H104875a.pdf

Chao, G. and M. Dyer, 2000. Word sense disambiguation of adjectives using probabilistic networks. Proceeding of the of the 18th Conference on Computational Linguistics, July 31-August 4, Saarbrucken, Germany, pp: 152-158. DOI: 10.3115/990820.990843.

Elsayad, A.M. 2010. Predicting the severity of breast masses with ensemble of Bayesian classifiers. J. Comput. Sci., 6: 576-584. ISSN: 1549-3636

Hoeven, V.D.G., A. Andernach, V.D.S. Burgt, G.J. Kruijff and A. Nijholt *et al*., 1996. SCHISMA: A natural language accessible theatre information and booking system. Proceeding of the of the 1st International Workshop on Applications of Natural Language to Data Bases, June 28-29, Versailles, France, pp: 271-285.

Hulstijn, J., 2000. Dialogue Models for Inquiry and Transaction. Ph.D. thesis, University of Twente, Netherlands.

Ian, W. and E. Frank, 2005. Data Mining: Practical Machine Learning Tools and Techniques. 2nd Edn., Morgan Kaufmann, San Francisco, ISBN: 10: 0-12-088407-0, pp: 560.

Intel, 2004. Probabilistic network library: User guide and reference manual. Intel Corporation. Implementation. https://sourceforge.net/projects/openpnl

Keizer, S. and O.P.R. Akker, 2007. Dialogue act recognition under uncertainty using Bayesian networks. Natural Language Eng., 13: 287-316. DOI: 10.1017/S1351324905004067

Keizer, S., 2003. Reasoning under uncertainty in natural language dialogue using Bayesian networks. Ph.D. thesis, University of Twente, Netherlands.

Khor, K.C., C.Y. Ting and S.P. Amnuaisuk, 2009. From feature selection to building of Bayesian classifiers: A network intrusion detection perspective. Am. J. Applied Sci., 6: 1949-1960. ISSN: 1546-9239

Lemon, O., P. Parikh and S. Peters, 2002. Probabilistic dialogue modeling. Proceeding of the of the 3rd SIGDIAL Workshop on Discourse and Dialogue, July 11-12, Philadelphia, Pennsylvania, pp: 125-128. DOI: 10.3115//1118121.1118138.

Mehdi, M., S. Zair, A. Anou and M. Bensebti, 2007. A Bayesian networks in intrusion detection systems. J. Comput. Sci., 3: 259-265. ISSN: 1549-3636

Murphy, K., 2002. Dynamic Bayesian networks: Representation, Inference and Learning. Ph.D. thesis, University of California, Berkeley.

Mustapha, A., M.N. Sulaiman, R. Mahmod and H. Selamat, 2008. Classification-and-ranking architecture for response generation based on intentions. Int. J. Comput. Sci. Network Security, 8: 253-258. http://paper.ijcsns.org/07_book/200812/20081236.pdf

Mustapha, A., M.N. Sulaiman, R. Mahmod and H. Selamat, 2009. A Bayesian approach to intention-based response generation. Eur. J. Sci. Res., 32: 477-489. ISSN: 1450-216X

Mustapha, A., M.N. Sulaiman, R. Mahmod and H. Selamat, 2010. Corpus-based analysis on cross-domain experiments in classification-and-ranking generation. J. Comput. Sci., 6: 1305-1312. ISSN: 1549-3636

Narayanan, S. and D. Jurafsky, 2002. A Bayesian model predicts human parse preference and reading time in human sentence processing. In: Advances in Neural Information Processing Systems, Dietterich, T.G., S.B. and Z. Ghahramani (Eds.). MIT Press pp: 59-65.

Pulman, S., 1996. Conversational games, belief revision and Bayesian networks. Proceeding of the of the 7th Computational Linguistics in the Netherlands Meeting, November 15, Netherlands, pp: 1-25. DOI: 10.1.1.46.1838.

Ramakrishnan, G., A. Jadhav, A. Joshi, S. Chakrabarti and P. Bhattacharyya, 2003. Question answering via Bayesian inference on lexical relations. Proceeding of the Association for Computational Linguistics, Workshop on Multilingual Summarization and Question Answering, July 11, Sapporo, Japan, pp: 1-10. DOI: 10.3115/ 1119312.1119313.

Ribeiro-Neto, B., I. Silva and R. Muntz, 2000. Bayesian network models for Information Retrieval. In: Soft Computing in Information Retrieval: Techniques and Application, Crestani, F. and G. Pasi (Eds.). Springer Verlag, pp: 259-291.

Saat, N.Z.M., K. Ibrahim and A.A. Jemain, 2010. Bayesian methods for ranking the severity of Apnea among patients. Am. J. Applied Sci., 7: 167-170. ISSN: 1546-9239

Zweig, G. and S. Russell, 1999. Probabilistic modeling with Bayesian networks for automatic speech recognition. Australian J. Intel. Inform. Process. Syst., 5: 253-60. http://research.microsoft.com/pubs/77974/ajiips-speech.pdf