# A BAYESIAN NETWORKS APPROACH FOR DIALOG MODELING: THE FUSION BN

*F. Fernandez Martinez, J. Ferreiros, R. Cordoba, J.M. Montero, R. San-Segundo, J.M. Pardo*

Speech Technology Group, Universidad Politécnica de Madrid, Madrid, Spain.
{efhes,jfl,cordoba,juancho,lapiz,pardo}@die.upm.es

## ABSTRACT

Bayesian Networks, BNs, are suitable for mixed-initiative dialog modeling allowing a more flexible and natural spoken interaction. This solution can be applied to identify the intention of the user considering the concepts extracted from the last utterance and the dialog context. Subsequently, in order to make a correct decision regarding how the dialog should continue, unnecessary, missing, wrong, optional and required concepts have to be detected according to the inferred goals. This information is useful to properly drive the dialog prompting for missing concepts, clarifying for wrong concepts, ignoring unnecessary concepts and retrieving those required and optional. This paper presents a novel BNs approach where a single BN is obtained from N goal-specific BNs through a fusion process. The new fusion BN enables a single concept analysis which is more consistent with the whole dialog context.

***Index Terms—*** dialog modeling, bayesian networks

## 1. INTRODUCTION

A dialog is a communicative process aimed to negotiate and, eventually, satisfy some objective. In this sense dialog modeling plays a fundamental role in helping users to reach their dialog goals efficiently when interacting with speech interfaces. Recent approaches [1][2][3] have shown a significant effort toward the goal of applying BNs to spoken dialog systems, SDS. BNs enable true mixed-initiative dialog modeling allowing a more flexible and natural spoken interaction. This paper presents new advances on the application of BNs to dialog modeling focusing on both the appropiate number and topology of the considered BNs. In this sense, we compare the typical N goal-specific BNs approach with a new one based on a fusion BN, both in terms of performance (i.e. dialog goal identification and conceptual validation capabilities) and complexity (i.e. computational cost).

## 2. SYSTEM DESCRIPTION

The solutions that we are presenting in this paper have been explored in the development of a conversational interface that allows users to control a commercial Hifi audio system using natural language sentences. The Hifi system is constituted by a compact disc (with a charger of three discs), two tapes and a radio receiver. This system can be normally controlled by an infrarred remote control. Instead, users are going to control the system using a microphone. This interface makes the translation of speech into the corresponding IrDA commands needed to perform the desired actions according to the user's intention. A detailed description of the system can be found in [1].

## 3. DATA COLLECTION

The database is made up of $463$ control sentences collected from different users. In each sentence the user addresses the system in order to perform some actions.

Each sentence has been semantically tagged. A concept dictionary has been defined by an expert trying to cover all the relevant semantic categories in the domain. The resulting concepts can be grouped into: "actions" to be performed over the system (e.g. to play), "parameters" that can be configured in the system (e.g. the volume), and their corresponding "values" (e.g. a number).

In summary, there are a total of 58 concepts comprising 22 actions, 16 parameters and 20 values. Additionally, each sentence has been also tagged with its corresponding dialog goals according to the user's intention. A set of $15$ goals has been defined according to the available functionality. Table 1 shows an example.

**Table 1**. *Database description: example sentence.*

| **U:** "Play the third track from the first cd and raise the volume." | |
|---|---|
| **Concepts** | **Dialog Goals** |
| STATE_ACTION=[play]<br>TRACK_VALUE=[3]<br>TRACK_PARAM=[track]<br>DISC_VALUE=[1]<br>DEVICE_VALUE=[cd]<br>DISC_PARAM=[cd]<br>VOLUME_ACTION=[+]<br>VOLUME_PARAM=[volume] | "device selection"<br>"playing parameters definition"<br>"source state modification"<br>"volume adjustment" |

## 4. DIALOG MANAGEMENT BASED ON BNS

The first task of the Dialog Manager (DM) module is to identify the intention (i.e. dialogue goals) of the user considering the last utterance together with the dialog context. Then, according to the inferred goals the DM has to make a decision regarding how the dialog should continue. Both tasks can be accomplished using BNs.

### 4.1. "Forward Inference"

BNs can be adopted to model the existing causal relation between the goals and the concepts [1][2]. Typically, both of them are assumed to be binary [3] (i.e. a concept is true or "present" only when it is observed in the sentence). Thus, from the whole set of available evidences, e.g. $E = \{C_1 = 0, C_2 = 1, ..., C_N = 1\}$ for $N$ defined concepts, a posterior probability $P(G_i = 1|E)$ can be obtained for each goal using the "Forward Inference" technique, FI [2][4].

Subsequently, a decision is made for each BN on the comparison of the posterior with a defined threshold, $\theta$. As a result of that comparison, one goal is "present" if the corresponding posterior is over the threshold; otherwise the goal is "absent".

**Table 2**. *Concept analysis used to drive the dialog.*

| | $P(C_j = 1 \mid E^*) < \theta$ | $P(C_j = 1 \mid E^*) \geq \theta$ |
|---|---|---|
| $C_j$ absent<br>($C_j = 0$) | $C_j$ **unnecessary**<br>(No action) | $C_j$ **missing**<br>(Prompt to request $C_j$) |
| $C_j$ present<br>($C_j = 1$) | $C_j$ **wrong**<br>(Prompt to clarify<br>or notify about $C_j$) | $C_j$ **required**<br>($C_j$ is stored<br>in the dialog memory) |

### 4.2. "Backward Inference"

After the FI process, and assuming the inferred results (i.e. those goals which were decided to be "present", $G_i = 1$) as new evidences, Bayesian inference can be applied again but this time aimed at the estimation of $P(C_j = 1|E^*)$ (the probability that each concept should be present) where $E^* = \{G_1 = 1, G_2 = 0, \ldots, G_M = 1, C_1 = 0, C_2 = 1, \ldots, C_N = 1\}$. This process is known as the "Backward Inference", BI, technique [2].

Making a similar binary decision on the value of $P(C_j = 1|E^*)$, it is possible to check whether that concept should be "present" (i.e. $P(C_j = 1|E^*) > \theta$) or not.

#### *4.2.1. Concept analysis*

The BI result can be compared with the actual occurrence of the concept enabling the classification presented in Table 2.

As a result of that analysis [2] every concept can be properly classified allowing the DM to perform a suitable action (a possible dialog proceeding strategy has been suggested below each result). For example, the system can drive the dialog prompting for the "missing" concepts.

The accuracy of this analysis, as well as a correct identification of the corresponding dialog goals, is of vital importance to ensure the appopiate behaviour of the SDS. For instance, a possible missclassification may occur when considering a "required" concept as "wrong". In that case, the system would probably try to correct or clarify a concept that actually is not erroneous but needeed to satisfy the inferred goals. From this revealing example it is clear that the resulting misbehavior from a wrong concept classification has a disastrous impact on dialog regarding consistency, naturalness and sucess.

## 5. BASELINE APPROACH: N GOAL SPECIFIC-BNS

The baseline approach [1][2][3] assumes the dialog goals to be conditionally independent. Therefore, each corresponding BN is considered independently for each goal (left side of Figure 1).

### 5.1. Concept selection for BN development

In order to decide on which concepts should be part of each BN model an Information Gain (IG) criterion has been used [3]. Using IG measures for a particular goal and each defined concept it is possible to sort the list of concepts for that specific goal. For compactness reasons, once the list is sorted, we simply select the top M concepts (i.e. the most representative) which add up to a certain percentage of the overall IG for each goal. Consequently, each BN model only includes the subset of concepts ($C_j$) with the strongest dependency of a particular goal ($G_i$) according to the considered % of IG. The conditional probabilities that quantify those dependencies are estimated tallying the counts from training data.

## 6. NEW APPROACH: THE FUSION BN

Approaching BI independently, just considering the local context for each goal and thus only a subset of the whole evidence set, could drive the DM to make wrong decisions. This is specially decisive as we are considering multiple goals scenarios where the user typically refers to several dialog goals simultaneously. If each goal and its corresponding BN are handled individually, then different results according to Table 2 are possible for a particular concept depending on which BN the BI process is applied to.

For example, assuming that both goals in Figure 1 have been positively inferred trough FI, two different results are available for $C_2$. Thus, $C_2$ could be "required" for $G_1$ but "wrong" for $G_2$ instead. Opposite results would drive the analysis to different classifications of the same concept.

On the contrary a better solution is possible just merging all the BNs into one (Figure 1). As a result of having a unique BN, a single result is obtained for every concept conjunctly dependent on all goal and concept evidences since only one BI process is performed over the new larger BN. All goals are solved jointly in a common scenario so that a concept could not be regarded differently for each particular goal. Consequently, the main difference between the former approach and the new one is that both FI and BI results are consistent with the whole dialog context.

### 6.1. The Fusion algorithm

When fusion happens, a new BN arise merging both original BNs through their common concepts. Starting from the formerly obtained N goal-specific BNs (i.e. baseline approach), we describe next the basis for this fusion method.

1. Begin with a set $F_{BN}$ of N BNs, one per each defined goal $G_i$ consisting of the goal itself and the top $M$ concepts, i.e. $S(G_i)$, needed for the considered % of IG.
2. For each distinct pair of BNs, BN($x$) and BN($y$), check whether fusion is possible, i.e. if $S(x) \cap S(y) \neq \emptyset$, in which case:
   (a) Merge both BNs into a new larger BN, labeled BN($z$) where $z = \{x, y\}$, as described before.
   (b) Remove BN($x$) and BN($y$) from $F_{BN}$ replacing them by the new BN (i.e. insert BN($z$) into $F_{BN}$).
3. Repeat 2 while there are still possible fusions left to do.

After applying the fusion algorithm to the N initial BNs, it could be possible to obtain not a single BN but several of them. Indeed, there could be some goals that the model considers conditionally independent as they do not have any concept in common. In anycase, the corresponding analysis to every obtained fusion BN would have previously mentioned drawbacks solved.

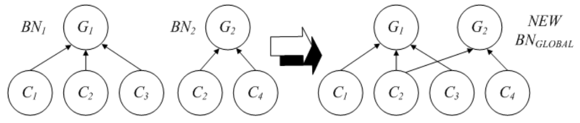

**Fig. 1**. *Detail of the BN merging.*

*6.1.1. Optimization of the BN model cost*

The fusion process always produces more complex BNs. Therefore, we need to ensure their trainability (thus avoiding sparsely trained BNs) and computational tractability. To that effect, a study on the optimum direction of the arcs that model the existing dependencies betweens goals and concepts can be done.

The independence assertions in a BN are important to reduce the complexity of inference [4]. On the other hand, the cost of the model, defined as the number of parameters of the corresponding conditional probability tables, may significantly vary according to the direction of the considered dependencies.

If we compute the cost of the resulting BN model from Figure 1 (details are presented below) for both possible options (each term of the sum is a power of two since we are assuming binary variables), we can conclude that "goal → concept" dependencies are better (and less costly) for this particular example. This result usually holds for BNs applied to DM. If the fusion process results in several BNs, this optimization should be done individually for each resulting BN.

$$Cost_{concept\rightarrow goal} = \underbrace{2^4}_{P(G_1|C_1,C_2,C_3)} + \underbrace{2^3}_{P(G_2|C_2,C_4)} +$$

$$+ \underbrace{2^1}_{P(C_1)} + \underbrace{2^1}_{P(C_2)} + \underbrace{2^1}_{P(C_3)} + \underbrace{2^1}_{P(C_4)} = 32$$

$$Cost_{goal\rightarrow concept} = \underbrace{2^1}_{P(G_1)} + \underbrace{2^1}_{P(G_2)} +$$

$$+ \underbrace{2^2}_{P(C_1|G_1)} + \underbrace{2^3}_{P(C_2|G_1,G_2)} + \underbrace{2^2}_{P(C_3|G_1)} + \underbrace{2^2}_{P(C_4|G_2)} = 24$$

## 7. EXPERIMENTAL SETUP

A stratified 10-fold cross-validation on the whole data set has been performed for all the experiments. Cost optimization has also been applied to baseline BN models. Table 3 shows the resulting cost for each tested solution.

For benchmarking purposes we considered first the 100% of IG ("ALL" model) and then compared it with lower percentages. In order to estimate the optimum threshold, several values have been tested from $0,1$ to $0,9$ using a $0,1$ step.

We have included the typical goal or topic classification performance measure, the "F-measure" [5] estimated as a function of the "recall", $R$, and "precision", $P$ measures [5]. We gave $R$ and $P$ equal importance (i.e. $\beta = 1$).

We have not used the FI results when performing the BI process (i.e. only tagged goals are part of the set of available evidences). Hence FI errors do not affect to BI performance since both processes have been evaluated independently.

All the sentences in the dataset are assumed to be "self-contained". Therefore, we assume that there is not any "missing" concept regarding the dialog goals that each sentence has been labelled with (i.e. all tagged concepts regarded as "required").

**Table 3**. *Estimated cost (# of params) for each solution.*

| IG% | Baseline | Fusion |
|---|---|---|
| ALL | 2670, 00 | 2.883.614, 04 |
| 90 | 1727, 20 | 787.924, 81 |
| 80 | 1284, 80 | 398.721, 21 |
| 70 | 940, 80 | 208.736, 80 |
| 60 | 672, 80 | 71.788, 80 |
| 50 | 461, 20 | 11.774, 20 |

## 8. EVALUATION RESULTS AND DISCUSSION

### 8.1. Fusion BN approach evaluation

Figures 2 and 3 shows respectively the FI and the BI results obtained for different IG%. As it can be observed in both figures, better results are obtained as the IG% decreases. This result could seem a bit surprising. Adding a greater number of concepts to a BN we are supposedly able to model more accurately the probabilistic dependencies between concepts and goals. Unfortunately, our models have become too complex by comparison to the amount of data available.

Both figures also illustrate classical $R$-$P$ trade-off. $R$ increases monotonically as the decision threshold increases, whereas $P$ decreases. This behavior can be intuitively deduced from the parabolic shape of the curves (max accuracy is well defined near $0,4$ threshold). Best FI and BI results are $88,14\%$ (50% IG model) and $81,00\%$ (60% IG model) respectively.

Keeping the $R$ and $P$ trade-off in mind, we decided to combine both FI and BI measures through a new F-measure estimation. Therefore, we used the FI and the BI results, respectively, in place of $R$ and $P$ and adopted $\beta = 1$ to ensure that we were maximizing both. Subsequently, following the typical early stopping method we reached a maximum combined performance of $84,09\%$ (60% IG model for a $0,5$ threshold).

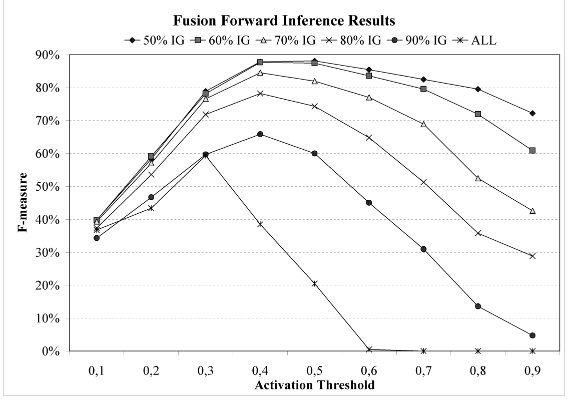

**Fig. 2**. *FI results for fusion approach.*

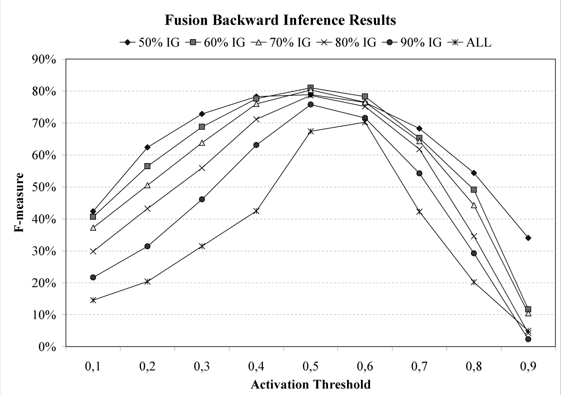

**Fig. 3**. *BI results for fusion approach.*

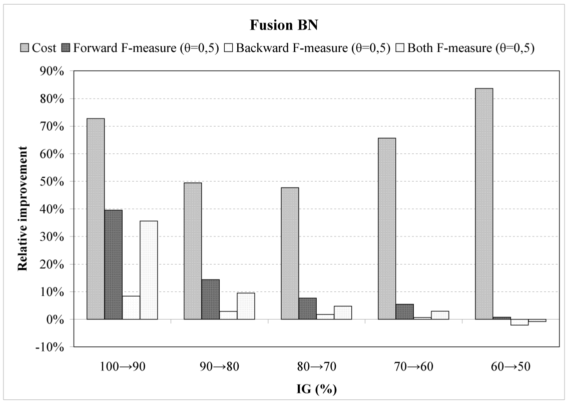

**Fig. 4**. *Combination of FI and BI results for fusion approach.*

Figure 4 shows, for a $0,5$ threshold, the relative variations when the % of IG changes as stated in each column set. In all cases a $10\%$ of absolute reduction of IG is performed. Relative variations regarding cost, FI and BI accuracies and also their combination are presented. Hence, the "$100 \rightarrow 90$" results correspond to the comparison between the "ALL" ($100\%$) and the $90\%$ of IG models. As it can be observed, the relative improvement regarding the FI and BI combination ("Both F-measure" in the figure) becomes negative (i.e. decreases a $0,84\%$) when considering a $50\%$ of IG thus defining our stop point. Nonetheless, the $50\%$ IG model is $83,60\%$ less costly than the $60\%$ IG model which could be consider a fair trade-off.

### 8.2. Baseline approach comparison

Finally, figures 5 and 6 respectively show the best FI and BI results obtained for both approaches.

Proceeding in the same way, similar experiments to those previously presented were performed for the baseline approach. However, in order to fairly compare both approaches, for each % of IG model we directly combined the best FI and BI results regardless of the threshold we used to obtain them. Consequently, the best combined result for the fusion approach could be slightly improved up to $84,22\%$, with a $60\%$ of IG model that performs $87,70\%$ FI accuracy considering a $0,4$ threshold and $81,00\%$ BI accuracy with a $0,5$ threshold. On the other hand, the best combined result for the baseline approach is obtained for the "ALL" model and rises up to $71,20\%$ by considering respectively $0,9$ and $0,3$ as the FI and BI thresholds (FI $92,29\%$ and BI $57,96\%$). The best fusion approach result ($84,22\%$) means a $13,02\%$ of combined accuracy improvement compared to the baseline approach.

As it can be observed in the Figures, the baseline approach outperforms the fusion approach regarding FI performance by a $4,59\%$. On the contrary, regarding BI the fusion approach clearly outperfoms the baseline by a substantial margin, a $23,04\%$ difference. That difference tips the balance in fusion's favor as it can be checked from both overall combined results.

However, as expected, the fusion approach is more costly than the baseline one. This is clear from costs reported in Table 3 although the complexity increase needed to outperform baseline models is not that important. As it can be deduced from the table, the least costly fusion model (50% IG) is "only" $4,4$ times more expensive than the best baseline model ("ALL") but its combined performance is still significantly better (i.e. $83,25\%$ obtained by combining $88,14\%$ and $78,87\%$, both by considering a $0,5$ threshold).

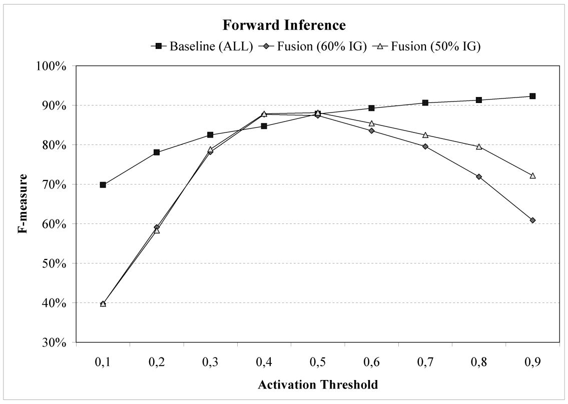

**Fig. 5**. *FI results comparison between both approaches.*

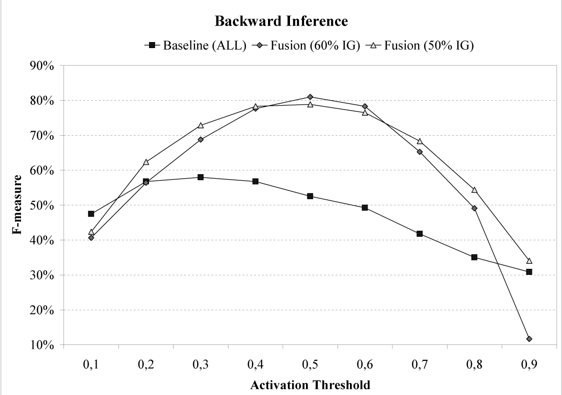

**Fig. 6**. *BI results comparison between both approaches.*

## 9. CONCLUSIONS

This work validates a new BN fusion model. The proposed fusion method provides a single BN that ensures that both the FI and the BI processes are performed within a whole common evidence context. The novel approach has been compared with a multiple goal-specific BNs based approach. Thanks to a much better concept classification (BI accuracy), the fusion approach results in a better overall (or combined) performance (roughly a $13\%$ better) in spite of its slightly worse goal classification (FI accuracy). New challenges arise as new approaches are possible, like using different models for FI and BI, or incorporating MDL techniques [3] to obtain enhanced fusion BNs.

## 10. REFERENCES

[1] F.Fernandez et al., "Speech interface for controlling an hi-fi audio system based on a bayesian belief networks approach for dialog modeling," in *Interspeech 2005*, Lisboa, Portugal, 2005.

[2] H. Meng, C. Wai, and R. Pieraccini, "The use of belief networks for mixed-initiative dialog modeling," in *IEEE Transactions on Speech and Audio Processing*, 2003, vol. 11, pp. pps. 757–773.

[3] H. Meng, W. Lam, and K.F. Low, "Learning belief networks for language understanding," in *Proc. of the ASRU*, 1999.

[4] C. Huang and A. Darwiche, "Inference in belief networks: A procedural guide," *I.J. Approximate Reasoning*, 1996.

[5] C.J. V.Rijsbergen, *Information Retrieval*, London, 1979.