Learning Belief Networks for Language Understanding

Helen M. Meng, Wai Lam and Kon Fan Low*

Human-Computer Communications Laboratory
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong, China
**Email: hmmeng@se.cuhk.edu.hk*

ABSTRACT

This paper is about learning Belief Networks (BNs) for spoken language understanding. The BNs are used to infer the communicative goal of a user's information-seeking query in a restricted domain. We assume that a restricted domain generally has a finite number of communicative goals. The problem is formulated as *N* binary classifications (one per goal), and each is performed by a BN. This formulation allows for the identification of queries with multiple goals, as well as queries with out-of-domain goals. The BN topologies are automatically learnt according to the Minimum Description Length (MDL) principle. We aim to learn the least complex topologies that can best model the available data set. These enhanced topologies are compared with a pre-defined, basic topology. Experiments with the ATIS-3 corpus shows that the enhanced topology improves goal identification accuracies from 83.7% to 91.5% when a single output goal is evaluated, and from 66.0% to 83.1% when multiple output goals are evaluated.

1. INTRODUCTION

This paper explores the use of Belief Networks (BN) for understanding spoken language. Information-seeking queries from a restricted domain often serves to convey a finite set of communicative goals. However, for a given communicative goal, the possible ways of expression are legion. BNs can be used to infer the communicative goal of the user from the query's semantics. Identification of the communicative goal helps to formulate a system's response most relevant to the user's query. This capability should be conducive towards the development of conversational systems that can respond to a user's query.

We believe that BNs offer several advantages to the problem of communicative goal identification. First, the dependencies between a query's communicative goal(s) and the relevant semantic concepts may be effectively captured by the topology of the BN. Second, BNs identify the communicative goal by means of probabilistic inferencing. This enables the use of data-driven approaches to alleviate the tedium in handcrafting heuristics. Third, BNs can handle situations where the input observations are incomplete, and thus may model spoken queries well. Fourth, the BN framework is suited for the optional incorporation of prior knowledge to aid inferencing.

Previously we have used BNs with a simple pre-defined structure (as shown in Figure 1) for the task of communicative goal identification. The user's query is first transformed into a sequence of semantic concepts. These are subsequently used by BNs to infer the query's communicative goal. The pre-defined BN structure models only the causal relationships between the goal and the concepts, while the concepts are assumed to be independent of one another. In this work, we attempt to use a more sophisticated topology for our BNs, which can model the inter-dependencies among the concepts. The new topology is automatically learnt from a training corpus. The learning algorithm tries to create the least complex network topologies which can most accurately model the training data. We expect the enhanced topology should lead to improved goal identification performance, without excessive demands on inference computation, or training corpus size.

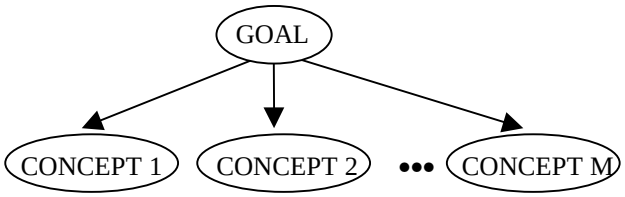

Figure 1. The pre-defined structure of our Belief Network. The arrows of the acyclic graph are drawn from cause to effect.

2. PREVIOUS APPROACHES

The use of BN for natural language processing has been attempted in [1]. Aside from using BNs, our problem has previously been tackled with alternative approaches. A well-known example is call-routing in AT&T's "How May I Help You?" task [2] and other similar call center tasks [3]. Here the caller's communicative goal determines the destination for call-routing. The problem is formulated as a topic identification or document classification problem, and for every input query the system outputs a single identified topic using a vector-based information retrieval technique. Other previous approaches include: grammar-based parsing [4] [5]; stochastic concept decoding with HMMs [6]; and probabilisitic recursive transition networks [7]. We have also applied a Naïve Bayesian approach previously [8].

3. TASK DOMAIN

Our experiments are based on the ATIS (Air Travel Information Systems) corpus [9]. We use the Class A sentences of ATIS-3, which have disjoint training and test sets of 1,564, 448 (1993 test) 444 (1994 test) transcribed utterances respectively. Each utterance (or query) is accompanied by its corresponding SQL query for retrieving the relevant information from the database. The main attribute label is treated as the communicative goal of

the query. There are 32 goals in the training set, but 11 of them covers over 95% of the training queries. We also found 43 training utterances with more than one goal. Examples include:

QUERY*: "chicago to san francisco on continental"*
GOAL*:* FLIGHT_ID

QUERY*: "give me the least expensive first class round trip ticket on u s air from cleveland to miami"*
GOALS*:* FLIGHT_ID, FARE_ID

The remaining attribute labels from the SQL are referenced as we enumerate the set of key semantic concepts for the ATIS domain.

4. PROCESS OF GOAL IDENTIFICATION

We have devised the following process for goal identification[10]:

1. *Semantic Tagging:* Each input query is tagged according to a set of semantic concepts. The tag sequence (sequence of semantic concepts) forms the input to our BNs, e.g.

   QUERY: "*may I have a listing of flight numbers from columbus ohio to minneapolis minnesota on Monday"*

   TAGS:<dummy><have><dummy><listing><prep> <flight_num><from><city_name><state_name><to> <city_name><state_name> <prep> <day_name>

   GOAL: FLIGHT_NUMBER

2. *BN development:* We develop one BN per communicative goal in the training corpus. We have 11 BNs in total to avoid using sparsely trained networks. The remaining goals are treated as out-of-domain (OOD). Each BN is trained to process an input query, and make a binary decision on the presence or absence of its goal. The pre-defined (basic) topology is as shown in Figure 1. For a network corresponding to goal G_i, we select the M concepts $\{C_1, C_2, \ldots C_M\}$ which have the highest Information Gain (IG) with G_i. These form the input of the BN. (Equation 1 is the formula for the IG between a concept C_k and the goal G_i.).

$$IG(C_k, G_i) = \sum_{c=0,1} \sum_{g=0,1} P(C_k = c, G_i = g) \log \frac{P(C_k = c, G_i = g)}{P(C_k = c)P(G_i = g)} \quad \ldots.(1)$$

   The selected concepts are regarded as most indicative of the goal. Each BN then applies Bayesian inferencing (Equation 2) to derive $P(G_i|\mathbf{C})$. This value is compared against the

$$P(G_i = 1|\vec{C}) = P(G_i = 1)\prod_{k=1}^{M} \frac{P(C_k = c_k|G_i = 1)}{P(C_k = c_k)} \ldots(2)$$

   threshold $\theta = 0.5$ to make the binary decision.

   Performance on binary classification is indicated by the F-measure (Equation 3), which combines both precision (P) and recall (R) for retrievial of goal G_i $(\beta=1)$.

3. *Goal Identification:* The decisions across all the BNs are combined to identify the output goal of an input query. We may label the query with the (single) goal giving the highest value of $P(G_i|\mathbf{C})$ across all BNs. Alternatively, we may label the query with all the goals for which the BNs voted positive – this achieves multiple-goal identification. In the case when all BNs vote negative, the input query is rejected as out-of-domain.

$$F = \frac{(1+\beta^2)RP}{\beta R + P} \ldots(3)$$

5. LEARNING NETWORK TOPOLOGIES

A Belief Network is a directed acyclic graph with nodes and arcs. The direction of the arc represents the probabilistic dependency between two nodes (or variables). The node where the arc arrives depends on the node where the arc originates. Hence the topology we have in Figure 1 assumes that the input concepts to the network are independent of one another. We wish to enhance this topology for inferring the goal, by capturing possible inter-dependencies amongst concepts. We believe this may improve the goal identification performance.

A machine learning technique is applied to learn the BN topology with reference to training data. Since it is computationally expensive to learn an arbitrary network, we constrain ourselves to topologies which belong to the classification-based network structures. A classification-based network has a root node (with no parents) which represents our goal G_i. The rationale is that the states of the class variables may depend on the membership of the goal, but not vice versa. Figure 2 shows an example of our enhanced topology. The root node for goal G_i only has arcs pointing outwards to the concept nodes, which is the same as the pre-defined topology. However, we now allow linkages between pairs of concepts, and the most desirable linkages will be learnt from training data.

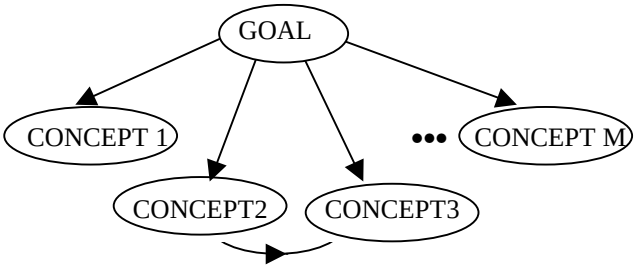

Figure 2. An example of a classification-based Belief Network, with an arc from concept node 2 to node 3.

The new topology will increase the complexity of our BNs. This should have two effects. On one hand, we can model the underlying distribution of the real data more accurately. On the other hand, the increased complexity will require more computation. We aim to learn network structures with minimal connectivity, but are sufficiently accurate for modeling the underlying distribution of the data. To achieve this we adopt Rissanen's minimum description length (MDL) principle [11].

5.1 The Minimum Description Length (MDL) Principle

The MDL principle offers a rigorous methodology for deriving a simple BN structure which is sufficiently accurate for modeling an available data set. Every node in the BN contributes towards the complexity of the network by a magnitude $L_{network}$ (the network description length). Lower values for $L_{network}$ reflect lower complexities. Each node also contributes towards the

accuracy in modeling data by a magnitude of L_{data} (the data description length). Lower values of L_{data} reflect higher accuracies. Consider the BN for goal G_j, with the root node $\boldsymbol{G_j}$, and concept nodes $\boldsymbol{\{C_{j1}, C_{j2}, \ldots C_{jM}\}}$. Also consider especially the concept node $\boldsymbol{C_{ji}}$ and its set of parent nodes P($\boldsymbol{C_{ji}}$). $L_{network}$ is defined as:

$$L_{network}(C_{ji}, P(C_{ji})) = k \log_2 N(s_i - 1) * \prod_{n \in P(C_{ji})} s_n +$$

$$(|P(C_{ji})|+1) \log_2 x \ldots\ldots (4)$$

where N is the total number of training utterances, s_i is the number of possible instantiations for $\boldsymbol{C_{ji}}$ (2 in our case), $|P(C_{ji})|$ is the number of parents for node C_{ji}, x is the number of nodes in the network, and s_m is the number of possible instantiations for a parent in the parent set. k is a constant for controlling the complexity of the network learnt.[1] L_{data} is defined as:

$$L_{data}(C_{ji}, P(C_{ji})) = \sum_{C_{ji}, P(C_{ji})} M(C_{ji}, P(C_{ji})) \log_2 \frac{M(P(C_{ji}))}{M(C_{ji}, P(C_{ji}))} \ldots\ldots (5)$$

where the summation is taken over all possible instantiations of the node $\boldsymbol{C_{ji}}$ and its parents. M(.) is the number of cases that match a particular instantiation in the training data.

The total description length (L_{total}) contributed by a given node, is defined in Equation 6.

$$L_{total}(C_{ji}, P(C_{ji})) = L_{network}(C_{ji}, P(C_{ji})) + L_{data}(C_{ji}, P(C_{ji})) \ldots\ldots (6)$$

5.2 Searching for the Topology with MDL

A best-first search algorithm is used to find the network topology of MDL. Our search procedure first computes the average L_{total} (between the 2 directions) for each arc that can be added to our predefined network structure.[2] The arcs are sorted in increasing order of L_{total} to form the sorted arc list, e.g. arc_1, arc_2… arc_n. Each arc in this list is paired with our pre-defined network structure (T_0) to form a search list of network-arc pairs, i.e. {(T_0 arc_1), (T_0 arc_2)… (T_0 arc_n)}. At this point iterative searching begins. Each network-arc pair receives an evaluation score, by summing the description length of the network and the average L_{total} of the arc. The search list is then sorted in increasing order of the evaluation score. The top network-arc pair is popped off the list, and the arc is inserted into the network to produce a new topology (T_1). The arc direction is selected to minimize the increase in description length. T_1 hence forms an enhanced topology of minimum description length, i.e. $T_{MDL} = T_1$. T_{MDL} is then paired with the arc_2 (which has the next lowest average L_{total}), and appended to the search list. The search list is sorted again and the process iterates to produce T_2. If T_2 has a lower description length than T_{MDL}, we set $T_{MDL} = T_2$. We ran our search procedure for a fixed number of iterations, and adopt the final T_{MDL} as our enhanced topology for the corresponding goal. Thus upon completion of learning, each goal has its own BN with an enhanced topology.

[1] The parameter setting is optimized with the 1994 test set, and we experimented with the 1993 test set.

[2] As mentioned previously, we only allow arc insertions between a pair of concept nodes in the BN.

6. EXPERIMENTS

As mentioned in section 3, our experiments are conducted using the ATIS-3 Class A sentences – training set and 1993 test set. Figure 3 compares the use of the pre-defined BN topology (Figure 1) with the automatically learnt topologies. Comparison is based on the F-measure obtained from training data. The figure shows that the enhanced topology led to significant performance improvement across the majority of our BNs

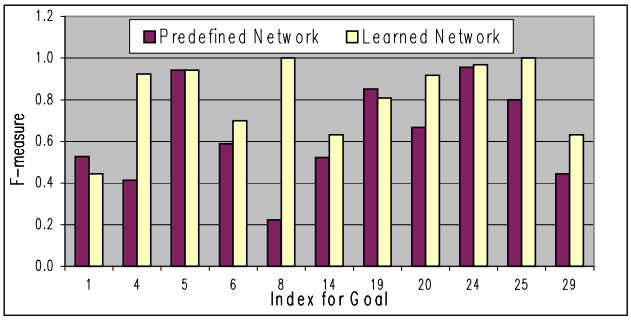

Figure 3. Comparison between the pre-defined topology and the automatically learnt topology, based on the F-measures of the 11 BNs applied to training data.

Next, we investigate to see if the enhanced topology contributes towards improvement in the overall goal identification accuracy. If all BNs vote negative for an input query, we reject it as OOD. We can adopt the *single-goal evaluation scheme* where we select G_k for the input query such that:

$$G_k = \arg\max_{j=1}^{M=11} P(G_j = 1 | C_j) \ldots\ldots (7)$$

In cases where the test query has multiple goals, identification is regarded as correct if G_k matches any one of the multiple goals. Results are tabulated in Table 1.

	Predefined Topology	Enhanced Topology
Total # test queries	448	448
Correctly Identified (multiple goals)	100% (8/8)	100% (8/8)
Correctly Identified (single goals)	88.9% (360/405)	94.6% (383/405)
Incorrect rejections	22	12
Correct rejection	20.0% (7/35)	54.3% (19/35)
Correctly handled	83.7% (375/448)	91.5% (410/448)

Table 1. Experimental results comparing the pre-defined and enhanced BN topologies based on overall goal identification performance (using a single output goal only).

We can also compared the two topologies based on *a multiple-goal evaluation scheme.* If one or more binary classifiers vote positive for the input query, all corresponding goals are

considered as identified goal(s) and are evaluated. Missing goals are treated as deletion errors and false positives are treated as insertion errors. Using the multiple-goal decision scheme, we found many insertion errors. Overall the pre-defined and enhanced topologies achieve goal identification performances of 66.0% (301/456) and 83.1% (379/456) respectively.

6.1 Analysis of Results

Figure 4 shows the topology learnt for goal 29 (ground_service.city_code). We see that linkages are added between the nodes FROM and TO, TRANSPORT_TYPE (e.g. "*limousine*") and TRANSPORT (e.g. "*ground transportation*").

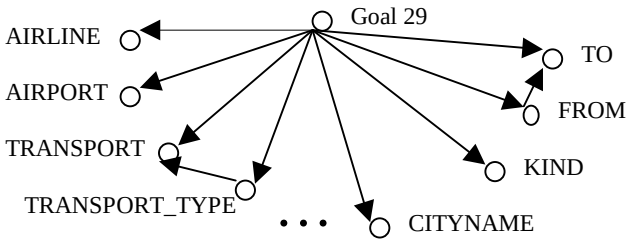

Figure 4. Enhanced topology automatically learnt for the BN for goal 29 (ground_service.city_code).

Capturing the dependencies between these concept pairs influences the probabilistic inferencing of the BN. It sharpens the output probabilities $P(G_i|\boldsymbol{C})$ to become more polarized around the threshold θ. As an illustration, if we sort the test queries for goal 4 in descending order of $P(G_4|\boldsymbol{C})$, then the ordered list for the pre-defined and enhanced toplogies are shown in Table 2. The letter in parentheses represents the truth, i.e. whether the test query belongs to goal 4 or not (Y for Yes and N for No).

Referring to Table 2, we see that when the pre-defined topology is used, our binary classifier (which thresholds at $\theta=0.5$) is bound to make mistakes with the second query, where $P(G_i|\boldsymbol{C})$=0.980, and the 16 instances where $P(G_4|\boldsymbol{C})$=0.845. However, for the enhanced topology, these probability values are polarized. There is a large gap between the values of 0.969 to 0.227. Investigation across our set of BNs shows that this polarization effect is instrumental in improving goal identification performance.

7. CONCLUSIONS AND FUTURE WORK

This paper reports our initial attempt in learning Belief networks for understanding natural language queries. The network topology is automatically learnt according to the Minimum Description Length principle, and serves to enhance our pre-defined topology by modeling the dependencies among semantic concepts in the query. Both topologies are compared for the task of communicative goal identification using the ATIS-3 corpus. Our results show that the enhanced topology brought about improvements in the F-measures for the majority of BNs in binary classification. Goal identification accuracies improved from 83.7% to 91.5% when we evaluate on a single output goal only. When multiple output goals are evaluated (and false positives penalized), accuracies improved from 66.0% to 83.1%. Future work include reducing false positives for performance improvement and investigating the robustness of the MDL methodology on speech recognition errors.

Goal 4 (airline.airline_code)
$P(G_4|C)$ for test queries, using the pre-defined topology:
0.991(Y), 0.980(N), 0.948(Y), 0.912(Y), 0.845(N), 0.845(N), 0.845(N)…0.741(Y), 0.664(Y), 0.529(Y), 0.046(N), 0.046(N)…
$P(G_4|C)$ for test queries, using the enhanced topology:
0.998(Y), 0.998(Y), 0.995(Y), 0.988(N), 0.985(Y), 0.985(Y), 0.969(Y), 0.227(N), 0.227(N)…

Table 2. Experimental results comparing the pre-defined and enhanced BN topologies, based on output probabilities $P(G_4|C)$.

8. REFERENCES

[1] Heckerman, D. and E. Horvitz, "Inferring Informational Goals from Free-Text Queries: A Bayesian Approach," Proceedings of the Fourteenth Conference on Uncertainty in Aritificial Intelligence, 1998, pp. 230-238.

[2] Arai, K., J. Wright, G. Riccardi and A. Gorin, "Grammar Fragment Acquisition using Syntactic and Semantic Clustering," Proceedings of the ICSLP 1998.

[3] Carpenter, B. and J. Chu-Carroll, "Natural Language Call Routing: A Robust, Self-Organizing Approach," Proceedings of the ICSLP 1998.

[4] Seneff, S., "TINA: A Natural Language System for Spoken Language Applications," Computational Lingusitics, Vol. 18, No. 1, 61-86, 1992.

[5] Ward, W. and S. Issar, "Recent Improvements in the CMU Spoken Language Understanding System," Proceedings of the ARPA Human Language Technology Workshop, 1994, pp. 213-216.

[6] Pieraccini, R., E. Tzoukermann, Z. Gorelov, J. Gauvain, E. Levin, C. Lee and J. Wilpon, "A Speech Understanding System Based on Statistical Representation of Semantics," Proceedings of ICASSP, 1992, pp. I-193 to I-196.

[7] Miller, S. and R. Bobrow, "Statistical Language Processing Using Hidden Understanding Models," Proceedings of the Human Language Technology Workshop, 1994, pp. 278-282.

[8] Meng, H., W. Lam, K. Low, " A Bayesian Approach for Understanding Information-seeking Queries," Proceedings of the International Conference on Systems, Man and Cybernetics, 1999, forthcoming.

[9] Price, P., "Evaluation of Spoken Language Systems: The ATIS Domain," Proceedings of the ARPA Human Language Technology Workshop, 1990, pp. 91-95.

[10] Meng, H., W. Lam and C. Wai, "To Believe is to Understand," Proceedings of Eurospeech, 1999.

[11] Rissanen, J., "Modeling by shortest data description, Automatica, vol. 14, pp. 465-471, 1978.