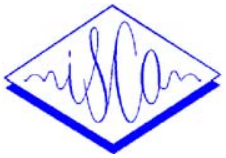


6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

TO BELIEVE IS TO UNDERSTAND

Helen M. Meng, Wai Lam and Carmen Wai*

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

10.21437/Eurospeech.1999-445

ABSTRACT

This paper is about language understanding using Belief Networks. Language understanding is a key technology in human-computer conversational systems. These systems often need to handle information-seeking queries from the user regarding a restricted domain. We devised a method for identifying the user's communicative goal(s) out of a finite set of within-domain goals. The problem is formulated as *N* binary decisions, each performed by a Belief Network. This formulation allows for the identification of queries with multiple goals, as well as queries with out-of-domain goals. Experiments with the ATIS corpus shows that around 90% of the user queries are correctly handled via goal classification, rejection or multiple goal identification.

Keywords: language understanding, belief networks

I. INTRODUCTION

One of the key technologies in a human-computer conversational system is spoken language understanding. State-of-the-art conversational systems can respond to the user's information-seeking queries for a restricted domain. These queries can often be classified into several domain-specific types. However, for a given query type, i.e. the communicative goal for a query, the possible ways of expression are legion. Understanding involves identifying the communicative goal from the query's semantics, and subsequently retrieving the relevant information to produce a coherent response.

As an example, we can consider an enormously simplified weather domain, which only consists of three semantic concepts: <weather>, <location> and <date>. A query which specifies two of the three concepts is likely to be asking for the missing one, e.g. "What do we expect for Hong Kong tomorrow?" A query containing all three concepts is likely to be asking for a yes/no response, e.g. "Will there be sunshine in Hong Kong tomorrow?" Another example is call-routing in AT&T's "How May I Help You?" task [1] and other similar call center tasks [2]. Here the caller's communicative goal determines the destination for call-routing.

Previous approaches to this problem include: (i) The use of heuristics to map a parsed query into an interpretation. The "parse" may be the output of a grammar-based parser [3], [4], or a stochastic concept decoder, e.g. HMMs [5], or probabilistic recursive transition networks [6]. Here the interpretation adopted depends primarily on an evaluation among parse alternatives, and sometimes the heuristics may not identify the best interpretation if task knowledge were to be considered. As a result, [4] had proposed a "beam of interpretations" approach, where multiple interpretations from multiple parses are used. (ii) The use of a vector-based information retrieval technique [2]. The problem is formulated as a topic identification or document classification problem, and for every input query the system outputs a single identified topic.

Our work bears resemblances to both streams. We use a semantic tagger to transform the input query into a sequence of semantic concepts. These form the input to our Belief Networks (also known as Bayesian Networks) for inferring the query's communicative goal. We believe that BNs offer several advantages to our problem [7]. First, the dependencies between a query's communicative goal(s) and the relevant semantic concepts may be effectively captured in the topology of the BN. Second, BNs identify the communicative goal by means of probabilistic inferencing. Under situations where massive data is involved, this provides an attractive alternative to handcrafting the heuristics between parses and their interpretations. 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 in order to aid the inference process.

II. TASK DOMAIN

We have chosen the ATIS (Air Travel Information System) domain [8] to investigate the feasibility of using BNs for language understanding. ATIS is a common task in the ARPA (Advanced Research Projects Agency) Speech and Language Program in the USA.

Our experiments are based on the Class A sentences of the ATIS-3 corpus, with disjoint training and test sets of 1,564, 448 (1993 test) 444 (1994 test) transcribed utterances respectively. Each utterance (or query) is accompanied with its corresponding SQL query for retrieving the relevant information. Thus we derive the communicative goal for each utterance from the main attribute label of its SQL query. In our training set, we counted a total of 32 communicative goals. We also found 43 training utterances with more than one communicative 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

III. SEMANTIC TAGGING

Semantic tagging abstracts the words in a query into a set of semantic concepts. While the main attribute label(s) in the SQL query is adopted as the communicative goal(s), the remaining attribute labels are identified as key semantic concepts for the ATIS domain, and served as a reference when we design our semantic tags for labeling an input transcription. We have a total of 60 hand-designed semantic tags. We have also devised an automatic procedure for

discovering such semantic categories from unannotated corpora [9].

Training utterances are automatically tagged, using a two-pass transformational procedure. This identifies the semantic concepts in the utterance transcriptions. The following shows an example of an utterance transcription and its corresponding tags:

QUERY: *what are the dinner flights from indianapolis to san diego on Wednesday may twelfth*

TAGS: <what><dummy><meal_description>
<flight><from><city_name><to>
<city_name><prep><day_name>
<month><day>

Henceforth each training query is represented by its annotated goal and a sequence of semantic concepts. These are used to train our BNs, as described in the following.

IV. *N* BINARY DECISIONS vs. ONE *N*-ARY DECISION

We need to infer an appropriate goal for a query, out of the finite set of goals in a restricted domain. One may formulate the problem as *N* binary decisions, or a single *N*-ary decision. We have chosen the former approach to facilitate the identification of cases with *multiple goals*, as well as the rejection of cases with *previously unseen, out-of-domain goals* [10]. We have also implemented the latter approach for benchmarking purposes.

Our approach utilizes multiple BNs – each a distinct classifier for making the binary decision regarding a unique goal. A BN outputs the confidence level for its decision regarding an input query, in terms of the *aposteriori* probability. One decision scheme is to adopt the goal with maximum *aposteriori* probability for the input query. With the use of a probability threshold, the BN output for a particular goal may be quantized into a binary decision. In this case we may utilize an alternative decision scheme: Queries for which all BNs vote negative are rejected as out-of-domain, and otherwise we revert to the maximum *aposteriori* rule.

V. THE BELIEF NETWORK

Each BN adopts a pre-defined structure as depicted in Figure 1. This simple structure models the causal relation between the concepts and the goal. Concepts within the query are assumed independent of each other.

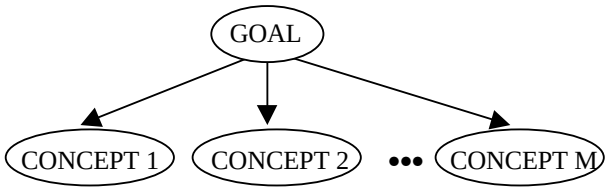


Figure 1 The pre-defined structure of our Belief Network. The arrows of the acyclic graph are drawn from cause to effect.

V.1 Concept Selection

For a given goal G_i, and its instantiations in the training set, we record the semantic concepts that are indicative of G_i. The recorded set is limited to M or below in size, in order to constrain computation during training. We compare the use of two measures to select the concepts with strongest dependency on G_i:

(i) Mutual Information, which measures the degree of co-occurrence *(i=0,1,2 N and k=1,2 M).*

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

$$MI(C_k, G_i) = P(C_k, G_i) \log \frac{P(C_k, G_i)}{P(C_k) P(G_i)} \quad (2)$$

(ii) Information Gain, which considers both the presence and absence of the concept and the goal.

Based on these measures, the top M semantic concepts will be selected as the features set for the *ith* goal, hence each goal may have a different set of selected concepts.

V.2 Bayesian Inferencing

Probabilities are estimated by tallying the counts from the training data. Each BN applies Bayes' Theorem:

$$P(G_i = 1 \mid C) = \frac{P(G_i = 1) P(C \mid G_i = 1)}{P(C)} \quad (3)$$

$$P(G_i = 1 \mid C) = \frac{P(G_i = 1) \prod_{k=1}^{M} P(C_k = c_k \mid G_i = 1)}{P(C)} \quad (4)$$

Two assumptions of marginal and conditional independence simplifies the above expression to:

V.3 Thresholding

As mentioned previously, the BN outputs its confidence level for the case that the input query is conveying its corresponding goal. Choosing a probability threshold allows for quantization of this confidence level into a binary decision. The threshold should be chosen such that we can maximize the performance on goal inference. Related performance measures include *recall (R)*, the percentage of queries correctly inferred by the BN for G_i out of all the G_i queries; and *precision (P)*, the percentage of queries correctly inferred by the BN for G_i out of all the inferred G_i queries. We combine both into a single score by optimizing with the *F-measure* [11]:

$$F = \frac{(1 + \beta^2) RP}{\beta R + P} \quad (5)$$

(β =1 in our experiments to treat precision and recall with equal importance).

VI. EXPERIMENTS

Inspection of the training utterances reveals that out of 32 goals, only 11 of them are instantiated 10 times or more. These 11 goals cover over 95% of the training set. Consequently, we have constructed 11 BNs, to avoid the use of sparsely trained BNs. The remaining goals and their utterances are treated as out-of-domain.

VI.1 Comparison between Mutual Information and Information Gain

For each of the 11 goals, we select (M=20) concepts with strongest dependency on the goal . *Mutual Information (MI)* and *Information Gain (IG)* are compared as the dependency measure. Only the selected goals in the training query are considered during classification, which maximizes the *aposteriori* probability according to Equation (4). [1]

Since IG considers both the presence and absence of concepts for goal classification, it can extract 20 concepts for all our goals. MI considers only the cases when a concept is present, and extracts fewer than 20 concepts for a number of goals. Therefore when MI is used we normalize the *aposteriori* probability prior to goal classification, by padding with a multiplicative constant of 0.5. Results of the comparison are tabulated in Table 1.

Concept Selection Measure	Performance (Training)
Mutual Information (MI)	85.42% (1336/1564)
Information Gain (IG)	93.67% (1465/1564)

Table 1. Comparison between the use of two different measures for concept selection (MI vs. IG), based on the goal classification accuracies on the training set.

It is observed that IG performs better than MI in concept selection for goal classification. This implies that the absence of certain concepts may be indicative of the communicative goal under some situations. To illustrate with an example, consider a query from our training set:

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>

REFERENCE GOAL: FLIGHT_NUMBER

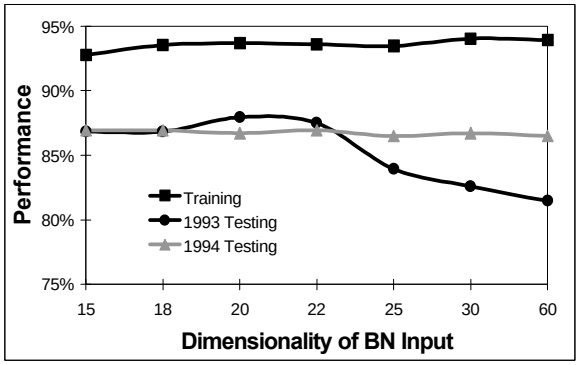
According to the set of concepts selected by MI for the goal FLIGHT_ID, all the query's semantic tags are indicative of the goal. The set of concepts selected by IG was similar, but it is also augmented by the *absence* of <flight_num>. The occurrence of <flight_num> in the input query lowered the *aposteriori* probability for FLIGHT_ID, which was eventually outweighed by FLIGHT_NUMBER.

VI.2 Varying the Input Dimensionality

A series of experiments were conducted in which we varied the BN input dimensionality, which is equivalent to the number of stored concepts per goal. Variation covered the range from 15 concepts to the full set of 60 concepts. The goal classification accuracies for the training set, the 1993 test set as well as the 1994 test sets are shown in Figure 2.

Performance accuracies in the plot are normalized based on the full size of the training / test sets. Hence queries which do not belong to the 11 goals are counted as errors. As observed in Figure 2, training accuracies increase with input dimensionality, while testing accuracies tend to decrease beyond 20 concepts per goal, possibly due to overfitting of the training data. This suggests that 20 concepts per goal is a suitable parameter setting. Performance with different normalizations are shown in Table 2.

Figure 2. Goal classification performance for different belief network input dimensionalities.



Normalization	Training	1993 Test	1994 Test
Over all queries	93.7% (1465/1564)	87.9% (395/448)	86.7% (385/444)
Queries of the 11 goals	97.1% (1465/1509)	95.4% (394/413)	94.59% (385/407)

Table 2. Goal classification accuracies computed using different normalizations – over the entire training/test sets, versus normalizing only over the relevant queries that belong to the 11 goals. 20 concepts per goal was used.

VI.3 Multiple Goals and Rejection

Thresholding enables the BN to make a binary decision about its goal. For a given query, we can look across all BNs to see if more than one network has voted positive (the case of multiple goals), or if all networks have voted negative (the case of unseen goal). A conversational system, which identifies an information-seeking query to have multiple communicative goals, may provide additional relevant information in the response. Alternatively if the query is identified with an unseen goal, it may be rejected as an out-of-domain query.

It may be reasonable to set the probability threshold at 0.5, since $P(G=1|C) + P(G=0|C)=1$. Alternatively we may also set the threshold at a value which maximizes the *F-measure*, as described in Section V.3. Comparative results are shown in Table 3.

Table 3 suggests that the threshold should be set by the F-measure, rather than at 0.5. This increases the rejection rate, but also improves the rejection accuracy. Moreover, it drastically reduces the number of queries identified to have multiple goals, while maintaining the same correct identification rate. Correctly handled cases include queries with correct goal classification, as well as those with correct rejection. Multiple goal queries are neither rewarded nor penalized. Comparison with our previous results (Table 3, bottom row) suggests that there is a slight performance advantage if an appropriate probability threshold is used.

VII. BENCHMARKS

As mentioned previously, to benchmark our approach we have also implemented the alternative approach using decision trees [12]. A single decision tree is grown to make the 11-way decision on goal classification. (without confidence levels in

[1] Since the number of concepts per goal is capped at M, our probability space does not sum to 1. However, we assume that it approximates 1 by using a large number of concepts with strong dependencies on the goal.

the output). The full set of 60 semantic concepts is used as input, and attribute selection for tree branching is based on the IG measure. Since the decision tree has no rejection capability, out-of-domain queries are counted as errors. Performance was 90.0% (403/448) for the 1993 test set, and 88.7% (394/444) for the 1994 test set. Comparison with the results in Table 3 (second last row) suggests that both approaches deliver comparable performance for our task.

We have also evaluated our outputs in terms of their extracted semantic category sequence [13]. This accounts for both concept and goal categories. Considering insertions and deletions, our error rates were 10.9% and 13.3% for the 1993 and 1994 test sets respectively.

	1993 Test		**1994 Test**	
Threshold	*0.5*	*F*	*0.5*	*F*
Classified √	389	386/413	378	377/407
# rejected	19	39	30	35
Rejection √	8/35	23/35	12/37	13/37
# multiple goals	119	51	69	22
Multiple goals √	5/8	5/8	4/6	4/6
Handled √	88.6% (397 of 448)	91.3% (409 of 448)	87.8% (390 of 444)	87.8% (390 of 444)
Handled √ (No Threshold)	87.9% (394/448)		86.7% (385/444)	

Table 3. Comparing the use of different probability thresholds – the use of 0.5 vs. other values which maximizes the F-measure. Comparison is based on test set performance – correct classification, number of rejected queries, correct rejection, number of multiple goal queries, correct multiple goal classification, and correctly handled queries overall.

VIII. CONCLUSIONS AND FUTURE WORK

This work is our initial attempt in applying Belief Networks for the identification of communicative goals in information-seeking queries. At present our BNs only model the causal relations between the query's semantic concepts and the underlying communicative goal. By formulating our *N*-way classification problem as *N* binary classifications, we are able to (i) identify queries with multiple communicative goals, and (ii) reject queries whose goals are outside of the prescribed knowledge domain, without significant loss in goal classification performance. Our experiments also found IG and the F-measure to be favorable, for their respective tasks of feature selection and probability thresholding in binary classifications. Future work includes experimentation with different BN topologies to capture inter-dependencies amongst concepts, as well as the incorporation of external knowledge (e.g. discourse) to improve goal inference performance.

REFERENCES:

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

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

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

[4] 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.

[5] 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.

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

[7] 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.

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

[9] Siu, K. C. and H. Meng, "Semi-automatic Acquistion of Domain-specific Semantic Structures," these proceedings.

[10] 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.

[11] van Rijsbergen, C. J., Information Retrieval. London. Butterworth. 1979.

[12] Quinlan, J. R., C4.5:Programs for Machine Learning, Morgan Kaufman, 1993.

[13] Minker, W., S. Bennacef and J. L. Gauvain, "A Stochastic Case Frame Approach for Natural Language Understanding," Proceedings of ICSLP 1996.