A Dynamic Semantic Model for Re-scoring Recognition Hypotheses

Carmen Wai, Roberto Pieraccini**, Helen M. Meng**
*Human-Computer Communications Laboratory
The Chinese University of Hong Kong
**SpeechWorks International
*{cmwai, hmmeng}@se.cuhk.edu.hk, **roberto.pieraccini@speechworks.com

ABSTRACT

This paper describes the use of Belief Networks (BNs) for dynamic semantic modeling within the United Airlines' FLight InFOrmation service (FLIFO). Callers can speak naturally to obtain status information about all flights (including arrival and departure times) of United Airlines. In this work we aim at enabling the application to utilize dynamic call information to improve speech recognition performance. Dynamic call information include the location of the caller, the time and date of the call, and the caller's dialog history. Dynamic semantic models can incorporate such additional information about the call in re-scoring the *N*-best recognition hypotheses. Our experiments showed that this improved the recognition accuracy of flight number utterances from 84.95% to 86.80%.

1. INTRODUCTION

This work strives to improve the recognition performance of a telephone-based spoken dialog system by incorporating dynamic information about the call or caller. *Dynamic semantic models* are applied to the United Airlines' Flight InFOrmation (FLIFO) service. *Dynamic semantics* refer to all the information we can gather about an incoming call, e.g. the location of the caller, the time and date of the call, or the caller's dialog history. It is conceivable that such information can contribute towards successful task completion in a spoken dialog system. For example callers are more likely to ask for information about flights arriving into or departing from the region where they are located. Dynamic semantics are drawn from different information sources, and vary from call to call. This is in contrast to the *static* information provided by the language model or pre-specified in the grammar for language understanding.

We have previously formulated a framework for natural language understanding and dialog modeling using Bayesian Belief Networks [1,2,3]. The current work extends the use of Belief Networks (BN) for our dynamic semantic models. For a given utterance, the BNs utilize dynamic call information to produce dynamic semantic scores, which are combined with the recognizer's confidence levels to re-score the *N*-best recognition hypotheses. We strive to improve overall recognition performance using this methodology.

2. THE FLIFO DOMAIN

We investigate the use of BNs for dynamic semantic modeling within the deployed United Airlines' FLight InFOrmation (FLIFO) service (1 800 824 6200). Callers are able to speak naturally to obtain information about thousands of United and United-related flights. They can check on the arrival and departure times as well as gate information even if they do not have the flight number. Table 1 shows a sample dialog in the FLIFO application.

Our experiments are based on the utterances of flight numbers only. Flight numbers may be verbalized in a variety of ways, e.g. 384 may be spoken as *three eight four*, *three hundred and eighty four*, *three eighty four*, etc. The recognizer's language model alone (e.g. bigrams) may not adequately capture the frequency information about specific semantic units (i.e. a specific flight numbers that are represented by several lexical forms) or their relation with other non-linguistic information. We use the dynamic semantic model to capture the relationships between a domain-specific semantic unit (e.g. flight number) and dynamic call information (e.g. the location of the caller).

We used disjoint training and test sets consisting of 2,837,861 and 1,485 flight number utterances respectively. These are obtained from real calls in the deployed system.

Each training utterance is associated with the top-scoring recognition hypothesis and its confidence level, as well as the following dynamic call information:

- Phone number (the first 3 digits being the area code)
- Time of the call
- Date of the call

The test set includes the *N*-best recognition hypotheses for each utterance, along with their confidence levels and the above dynamic call information.

System:	Welcome to the United Airlines flight information line. I can give you up-to-the-minute arrival and departure information for all United, United Express, Code Share and United Shuttle flights. Please say the United flight number, or say "I don't know it" and we'll get the flight number a different way.
User:	Flight one oh one
System:	Would you like the arrival or departure information for that flight?
User:	Arrival
System:	Okay, I'll look up flights which have that itinerary. Flight one oh one ….

Table 1. Sample dialog from the FLIFO application.

3. THE USE OF BELIEF NETWORKS FOR DYNAMIC SEMANTIC MODELS

3.1 Belief Network

A Belief Network (BN) is a graphical model representing probabilistic causal relationships among its nodes. In our implementation we used a pre-defined topology as shown in Figure 1. The arrows indicate the causal relationship between the goal and the concepts. This topology assumes that the concepts are independent of one another.

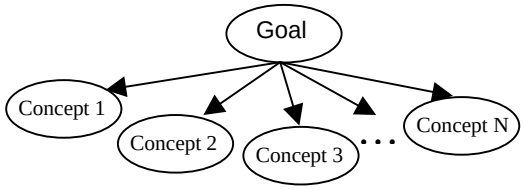

Figure 1. The pre-defined topology of our BNs. The arrows of the acyclic graph are drawn from cause to effect.

A dynamic semantic model (a BN as illustrated in Figure 1) is developed for each flight number. Each dynamic semantic model captures the relationship between a flight number and the following dynamic call information:

1. Area code of the caller's location. (We have categorized all area codes in the US into 57 regions[1])
2. Month in which the call was placed
3. Day of week in which the call was placed
4. Time of the call, categorized into 4 periods – morning, afternoon, evening and night.[2]

In order to further simplify the dynamic semantic models, each concept node receives a binary input. In total, each BN has 80 binary inputs: 57 locations, 12 months, 7 days of the week and 4 periods of call. The final topology is illustrated in Figure 2.

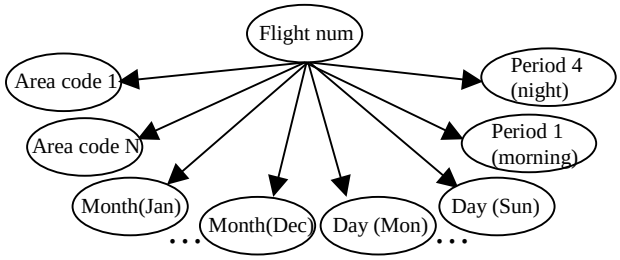

Figure 2. The topology of a dynamic semantic model (a BN) in the FLIFO domain.

Our corpus contains several million training utterances, but these have not been manually transcribed. We gathered the top ranking recognition hypothesis for each training utterance, and used only those with a confidence level above 800.[3] This subset is *assumed* to have correct transcriptions from recognition, and it is used to train our dynamic semantic models.

Further inspection of the training utterances showed that the training corpus includes 9706 distinct flight numbers, 3235 of which are instantiated in more than 100 utterances. Consequently, we developed 3235 dynamic semantic models (or BNs), one for each of the frequently inquired flight numbers. This avoids the development of sparsely trained BNs.

3.2 Bayesian Inferencing

As mentioned previously, each test utterance is associated with its *N*-best recognition hypotheses (in our case, $N = 2$) and their confidence levels. We re-score these hypotheses with the *dynamic semantic scores*, which are the output probabilities from the dynamic semantic models.

[1] http://thelist.internet.com/areacode.html

[2] Time of call is categorized as : 7-12:00noon (morning), 13-18:00 (afternoon), 19-24:00 (evening), 1-6:00 (night)

[3] Confidence levels from recognition ranges from 0 to 999.

Given an incoming call and a flight number utterance, the two-best recognition hypotheses correspond to two flight numbers and their associated dynamic semantic models. The dynamic call information is fed into each of the trained dynamic semantic model whose output dynamic semantic score is computed according to Equation (1). The score is the aposteriori probability that the flight number in question was actually uttered, conditional to the observed dynamic call information.

$$P(F_i = 1 \mid \vec{C}) = P(F_i = 1) \prod_{k=1}^{M} \frac{P(C_k = c_k \mid G_i = 1)}{P(C_k = c_k)} \qquad (1)$$

where F_i represents flight number i
C_k is the k^{th} concept (a binary input to the model)
M is total number of concepts (in our experiment M=80)

3.3 Flight Number Identification

For each recognition hypotheses, we combine the dynamic semantic score with the recognition confidence level in order to obtain a global score (*GS*), as defined in Equation (2):

$$GS_i = \lambda CL_i + (1-\lambda) B_i \qquad (2)$$

where i indexes the recognition N-best hypotheses (i =1 or 2)
GS_i is the global score for hypothesis i
CL_i is the recognition confidence level for hypothesis i
B_i is the dynamic semantic score for hypothesis i

We held out 1000 utterances from our test set to form a development test set for optimizing the free parameter λ. The remaining 485 utterances are reserved for evaluation.

The *N*-best recognition hypotheses are then re-ranked using their global scores. Our dialog system labels the test utterance with the flight number for which the global score is the highest. The entire procedure is illustrated in Figure 3.

4. RESULTS

The optimized value for λ is 2.65 x 10^{-5}, based on the development test set. The value is relatively small as it serves to balance the recognition confidence levels which have a large range (between 0 and 999), and the dynamic semantic scores which range only between 0 and 1. Results on the development and evaluation test sets are shown in Table 2. The performance on flight number identification is improved by 1.4% and 1.9% (absolute) in the development and evaluation test sets respectively.

	Flight number identification accuracies	
Test Set	**Hypothesis with max. recognition confidence level**	**Hypothesis with maximum global score**
Development (1000 utterances)	87.20%	88.60%
Evaluation (485 utterances)	84.95%	86.80%

Table 2. Performance on flight number identification in the United Airlines' FLIFO service. The center column shows accuracies on flight number identification based on the top-scoring recognition hypotheses. The right column shows accuracies on flight number identification based on re-ranking the *N*-best recognition hypotheses with dynamic semantic scores.

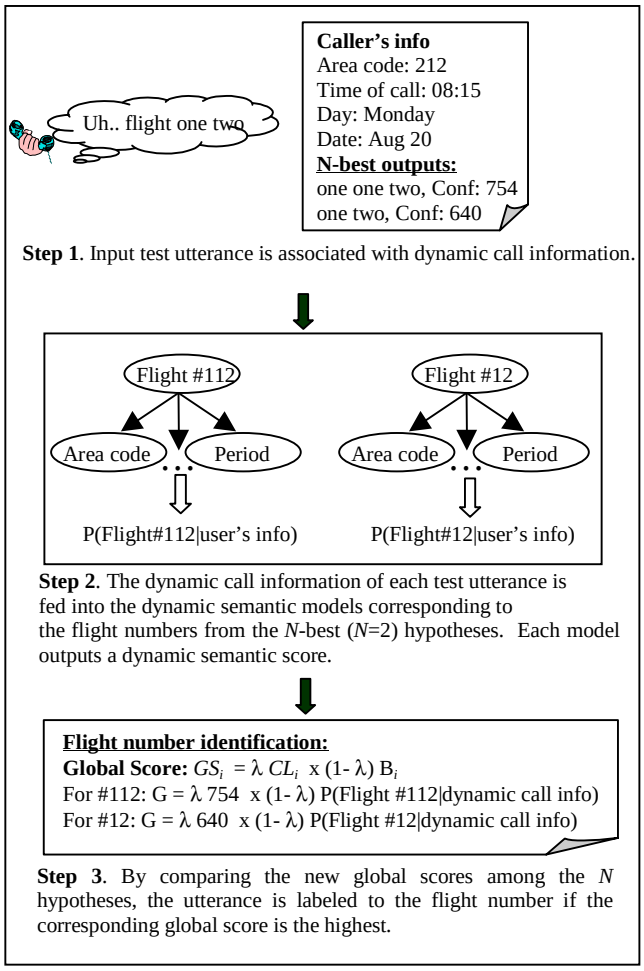

Figure 3. The use of dynamic semantic scores to re-rank the *N*-best recognition hypotheses (*N*=2).

Figure 4 shows the flight number identification accuracies resulting from maximizing (1) recognition confidence levels only; (2) dynamic semantic scores only; and (3) global scores combining the previous two elements. We observe that using the combined scores outperforms either score type alone.

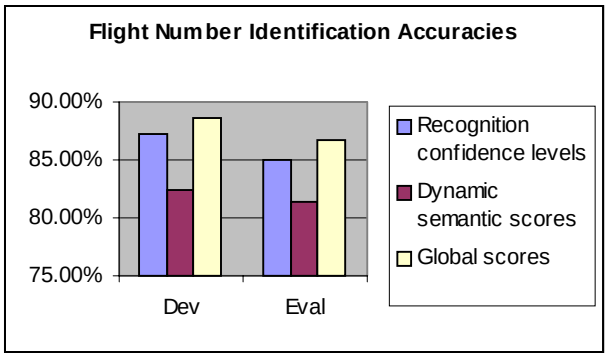

Figure 4. Flight number identification accuracies based on maximizing (1) recognition confidence levels only (left bar); (2) dynamic semantic scores only (center bar); (3) global scores which combine the previous two score types (right bar).

4.1 Significance Testing

Out of the 485 test utterances in our evaluation set, 412 (84.95%) flight numbers were identified correctly by maximizing the recognition confidence levels only; and 421 (86.80%) flight numbers were identified correctly by maximizing the global scores which incorporate dynamic call information. We conducted a significance test on the performance difference as follows:

The null hypothesis (H_0) and alternate hypothesis (H_1)are:

$H_0 : d = p_1 - p_2 = 0$

$H_1 : d = p_1 - p_2 < 0$

where $\alpha = 0.05$ *(significance level), hence* $Z_{0.05} = 0.8289$

p_1 *: proportion of correctly identified flight numbers by maximizing recognition confidence levels only*

p_2*: proportion of correctly identified flight numbers by maximizing the global scores*

$$Z_0 = \frac{p_1 - p_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

The test statistic is:

where $p_1 = 412/485$, $p_2 = 421/485$, $n_1 = n_2 = 485$ and

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{412 + 421}{485 + 485} = 0.8588$$

We reject H_0 if $Z_0 > Z_{0.05}$ or $Z_0 < -Z_{0.05}$.

Since $Z_0 = -0.8297 < -Z_{0.05}$, we conclude that using the global score can give better flight number identification accuracies than using the recognition confidence levels, and the performance difference is statistically significant.

5. RELATIVE IMPORTANCE OF THE DYNAMIC CONCEPTS

We further investigated the relative importance of the dynamic concepts – area code, month, day of week and time of call. We repeated our experiments by removing a single dynamic concept at a time. This entailed four sets of experiments where we removed, in turn, (1) the area code; (2) the month; (3) the day of the week; and (4) the time of the day. Concepts are "removed" by not providing any binary evidence for BN inferencing.

Table 3 shows the results of the four experiments based on the evaluation test set. Figure 5 illustrates the comparison between these experiments and the original setup with no dynamic concepts removed. It is observed that the greatest performance degradation (1.85%) in flight number identification occurred when the dynamic concept <area code> is omitted. This suggests that <area code> may be the most important concept among the various the dynamic call information. Removal of the concept <month> did not help performance, which suggests that flight inquiry is not affected seasonally.

Removed Concept	Flight number identification performance, based on maximizing the global scores
Area code	84.95%
Month	87.21%
Day of week	86.39%
Time of day	85.67%

Table 3. Performance accuracies in flight number identification upon removal of various dynamic concepts.

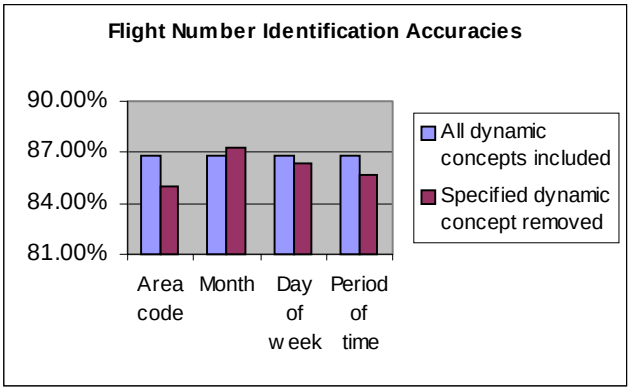

Figure 5. Flight number identification accuracies with (1) all dynamic concepts included (left bar); and (2) the specified dynamic concept removed (right bar).

5.1 Error analysis

We conclude that flight number recognition can benefit especially from knowledge of the caller's location. Table 4 illustrates how a top-ranking recognition hypothesis (based on recognition confidence levels) was a misrecognition initially, but re-ranking with dynamic semantic scores (especially the <area code>) recovered the correct hypothesis from rank=2.

Test utterance: forty nine, **Origin of call: Region 14 (derived from area code)**	
***N*-best results**: (1) Twenty nine, confidence level: 926 (2) Forty nine, confidence level: 806 Comment: Note that the second-best hypothesis is correct	
N.B. United Flight 29 flies from Philadelphia (Region 45) to Los Angeles (Region 7) United Flight 49 flies from San Francisco (Region 7) to Kahului (Region 14)	
Results with the global score (all dynamic concepts included) – Dynamic semantic score for (#29): 0.0006 – Dynamic semantic score for (#49): 0.0720 – Global Score (#29): 0.0251 – Global Score (#49): 0.0933 **Re-scored hypothesis: forty nine (Correct)**	**Results with the global score (dynamic concept <area code> removed)** – Dynamic semantic score for (#29): 0.0029 – Dynamic semantic score for (#49): 0.0025 – Global Score (#29): 0.0271 – Global Score (#49): 0. 0243 **Re-scored hypothesis: twenty nine (Wrong)**

Table 4. Example illustrating how dynamic call information (specifically, the area code) helped recover the correct recognition hypothesis initially ranked second in the *N*-best list. The caller is located in region 14, and is more likely to ask for about the flight number 49 which arrives at that region.

6. CONCLUSIONS

This paper describes our initial attempt in using Belief Networks as dynamic semantic models in the United Airlines' FLIFO domain. Dynamic semantic models are designed to capture the relationship among the flight numbers and dynamic call information, e.g. area code, date, and time of the call. By means of Bayesian inferencing with dynamic call information, our BNs output dynamic semantic scores, which are combined with recognition confidence levels to produce a global score. Re-ranking the *N*-best recognition hypotheses based on the global score shows improvement in flight number identification accuracies from 84.95% to 86.80%. The improvement was statistically significant. We have also observed that among the various dynamic concepts, <area code> is most important for flight number identification while <month> is least helpful.

ACKNOWLEDGMENTS

The authors wish to thank Drew Halberstadt, Zhihong Hu and Kristin Maczko from SpeechWorks International for their help.

REFERENCES

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

[2] Meng, H., W. Lam and K. F. Low, "Learning Belief Networks for Language Understanding," Proceedings of ASRU, 1999.

[3] Meng, H., C. Wai and R. Pieraccini, "The Use of Belief Networks for Mixed-Initiative Dialog Modeling," Proceedings of ICSLP, 2000.