Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 327-332 recibido 06-05-2005; aceptado 01-06-2005

Topic Identification based on Bayesian Belief Networks in the context of an Air Traffic Control Task

F. Fernández, L. F. D'Haro, J. Ferreiros, J. M. Montero, R. San-Segundo
Grupo Tecnología del Habla, Universidad Politécnica Madrid
Ciudad Universitaria, s/n, Madrid, Spain, 28040
{efhes, lfdharo, jfl, juancho, lapiz}@die.upm.es

Resumen: En este artículo presentamos una tarea de identificación de tópico basada en Redes Bayesianas. Estas redes son entrenadas a partir de los conceptos semánticos que se han etiquetado para cada frase a procesar y que han sido definidos por un experto en el dominio de aplicación. Los tópicos a identificar se corresponden con las cinco posiciones de control disponibles en un aeropuerto. Se ha llevado a cabo una evaluación basada en bloques de frases. Obtenemos una tasa de error de identificación de bloque del 3.5% para un esquema de evaluación 'winner takes all' usando un tamaño de 5 frases por bloque. Finalmente, comparamos los resultados obtenidos con una estrategia basada en un clasificador Bayesiano para el que tomamos como vector de parámetros las perplejidades resultantes de aplicar un modelo de lenguaje de tipo trigrama específico para cada uno de los tópicos. Los resultados obtenidos demuestran la importancia de considerar el orden de aparición de la información y la necesidad de incluirla en las Redes Bayesianas en futuros trabajos.
Palabras clave: Identificación de Tópico, Redes Bayesianas, N-gram, Control Tráfico Aéreo.

Abstract: In this paper we present a topic identification task based on a Bayesian Belief Network approach. These networks are trained with a number of semantic concepts which have been tagged for each utterance and defined by an expert in the application domain. The target topics are the five control positions available at the airport. In order to evaluate the performance of our approach we apply a block based evaluation scheme. The lower error rate that we obtained was 3.5% using a winner takes all evaluation scheme and using five utterances per block. Finally, we compare these results with those obtained by a Bayesian classifier considering a parameter vector constituted by the resultant perplexities, at phrase level, applying a trigram language model for each topic of the task; the obtained results allow us to know intuitively the importance of including temporal information into the BN in future works.
Keywords: Topic Identification, Bayesian Belief Networks, N-gram, Air Traffic Control

1 Introduction

Understanding natural language sentences for a specific application domain involves parsing the input sentence into a series of domain-specific concepts. These concepts correspond to those pieces of information that are relevant for the application. From the query's semantics it is possible to identify the communicative goal for the given query. Traditionally, several information retrieval techniques have been approached in order to map those queries into an interpretation e.g. the use of heuristics coded into a set of handcrafted context dependent rules (Ferreiros et al., 1998), or a stochastic concept decoder e.g. HMMs (Pieraccini et al., 1992). In this paper we are going to formulate the problem as a topic identification or document classification problem. Thus for each query we are going to infer the most likely corresponding topic.

The Bayesian Belief Networks (BN) have been previously used for language understanding (Heckerman and Horvitz, 1998). Moreover, BN have been applied in the context of dialogue modelling where first the BN are used to infer the user's informational goal (Meng et al., 2003) and subsequently to automatically detect, using the backward inference technique based on the inferred goal, missing or spurious concepts which the system has to request or clarify with the user and

ISSN: 1135-5948

triggering the appropriate system's response. Other previous approaches using BN for topic or goal identification were attempted in (Meng et al., 1999a; Meng et al., 1999b) in the context of the ATIS task.

We will identify the topic, or control position, corresponding to every processed utterance using the semantic concepts, tagged by an expert in the application domain, for each controller's utterance. With those concepts we will form an input to the BN models in order to infer the corresponding topic by means of Bayesian Inference. Thus, dependencies between a query's communicative goal(s) and the relevant semantic concepts are going to be effectively captured in the topology of the BN.

This paper is organized as follow: in section 2 and 3 we will describe the task domain and database setup. In section 4 and 5, the selection of the concepts and topic identification procedure are described. In section 6 we will show the experiments we carried on using the BN and finally, in section 7, a comparison approach using a Bayesian classifier.

2 Task domain: INVOCA project

The "INVOCA" project, (INVOCA, 2002), was done by the GTH. The main objective of this project is the detection of the relevant semantic frames present in the utterances spoken by the controllers through the communication channels.

Conversations between a pilot and a controller at the control tower could take place as much in English as in Spanish. A language identification module was used in order to distinguish between both possibilities. Those dialogues are very restricted since the speakers are forced to use a standard phraseology which is ruled by a set of syntactic-semantic constraints. Most of the times the controller's role is to provide a set of information items to the pilot while the pilot simply reads back the data received. Consequently, the dialogue is reduced to those communications in which one of the speakers needs to correct some miss-understanding data. Few utterances are needed to complete the process.

The functionality of the developed system included the five different air control positions available at the Madrid Barajas international airport: *"Arrivals"*, "*Clearances"*, *"Takeoffs"*, *"North Taxing"* and *"South Taxing"*.

In order to recover the relevant semantic frames and its values, the output of a continuous speech recognizer module was processed by a language understanding module that was constituted by a set of handcrafted context dependent rules.

3 Database setup

The database is made up of phrases from controller speakers corresponding to the different topics and languages. Each phrase is tagged by hand according to the specific set of semantic concepts defined by a linguistic and an expert in the application domain. Next we show an example:

QUERY: *Airnostrum eight seven six eight, wind three one zero seven knots cleared for take off runway three six left.*
CONCEPTS:
call_sign = [*airnostrum8768*]
takeoff_clearance = [*cleared take off*]
wind_info = [*310*]
wind_speed = [*7*]
runway = [*36left*]
TOPIC: *"Takeoff"*

Table 1 shows the existing distinction between examples in English and examples in Spanish of the original database. Actually we will not make any distinction about the language since, at a semantic level, tags are common to both languages. Besides, it is evident that the database has a significant imbalance. There are much more examples for the *"Clearances"* position. The main reason is that this position concentrates the highest density of traffic at the airport so much more utterances were recorded and processed.

		SPA	ENG	ALL
TRAIN	**ARRIVALS**	136	34	170
	CLEARANCES	3160	312	3472
	TAKEOFFS	141	47	188
	NORTH TAXING	145	30	175
	SOUTH TAXING	97	45	142
	ALL	3679	468	**4147**
TEST	**ARRIVALS**	58	15	73
	CLEARANCES	1353	134	1487
	TAKEOFFS	60	20	80
	NORTH TAXING	62	14	76
	SOUTH TAXING	42	20	62
	ALL	1575	203	**1778**

Table 1: Database Setup.

4 Concept selection for BN development

A BN is a directed acyclic graph with nodes and arcs where the direction of the arcs represents the probabilistic dependency between two nodes. The arrows of the acyclic graph are drawn from cause to effect. Assuming the structure depicted in Figure 1 we are modelling the causal relation between the topic and the concepts. We are going to develop one BN per topic.

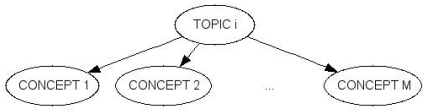

Figure 1: Basic structure of a BN.

The concepts we used were directly those handled by the semantic parser of the original system. There are a total of 60 concepts that have been defined according to a relevance criterion of each concept for the control task. However, in order to constrain the computational cost and to avoid sparsely trained models we should select a subset of concepts for each topic that are indicative of it.

We have used an Information Gain (IG) measure (see Equation 1) to select the concepts (C_k) with the strongest dependency of topics (T_i). The advantage of this measure respect to others e.g. Mutual Information, MI, is that it takes into account both the presence and the absence of each concept and topic (Meng et al., 1999a). With this information we create a sort list of concepts from where we select the top M concepts which sum up to a certain percentage of the overall IG for each topic.

$$IG(C_k, T_i) = \sum_{c=0,1}\sum_{t=0,1} P(C_k = c, T_i = t)\log\frac{P(C_k = c, T_i = t)}{P(C_k = c)P(T_i = t)}$$

Equation 1: Information Gain.

For benchmarking purposes we consider first the 100% of IG and then compare it with lower percentages alternatives. If we consider the 100% of IG for each topic we would have to select all the concepts as the inputs for each BN corresponding to a particular topic. However this solution would not be suitable because it would increase greatly the amount of necessary data to train the big networks that it would form. Consequently we must assume a reasonable limit for the maximum number of concepts that can be used as inputs for the BN.

We have defined a maximum of 15 concepts per BN. Anyway, this limit supposes almost the 100% of IG (i.e. the contribution from the rest of 45 concepts is not significant). We have also evaluated for 80% and 90% of IG.

5 Topic identification

Each BN is defined by a specific goal T_i and a set of input concepts C_k. We have assumed that the topics and the concepts are all binary, so the concept C_k is true (C_k =1) when it is observed in the sentence, that is, if it is included in the list of concepts tagged for that sentence.

We adopt a single goal or a winner takes all evaluation scheme since we assume that each utterance is specific of particular topic. Thus after applying Bayesian Inference (see Equation 2, Bayes' Theorem assuming marginal and conditional independence, which is equivalent to a naïve Bayes' formulation; M is the number of input evidences for the BN corresponding to topic T_i) the BN with the maximum a posteriori probability for the current input is considered as the identified topic. Probabilities are estimated tallying the counts from training data. Details regarding the inference algorithm we have used can be checked in Huang and Darwiche (1996).

$$P(T_i = 1|\boldsymbol{C}) = P(T_i = 1)\prod_{k=1}^{M}\frac{P(C_k = c_k|T_i = 1)}{P(C_k = c_k)}$$
$$where\ \boldsymbol{C} = \{C_1 = c_1, C_1 = c_2, ..., C_M = c_M\}$$

Equation 2: Bayesian Inference.

6 Experiments and Results

Next we present the topic identification results obtained for the different information gain percentages. In order to study in detail the possibilities that the BN offer, a block based evaluation scheme has been used.

A block is defined as set of consecutives phrases from which we are going to carry out topic identification. We want to study the effect on the topic identification of the number of phrases the block consists of. We believe as the blocks become greater, more evidences are available to identify the topic.

The building of the different blocks has been carried out following a window procedure. As a result we introduce one phrase shift between two consecutive blocks. This allows us to evaluate with more data and always with the same number of blocks.

6.1 Definition of the topic set

Initially we assumed the five different control positions as our set of topics to be identified. However, preliminary experiments showed a significant confusability between the *"North"* and *"South Taxing"* topics. Certainly, if we consider the five topics defined set and check the following confusion matrix (see Table 2 obtained for a block size of five utterances and for BN models considering a 90% IG), we can extract an interesting conclusion.

	Others	North T.	South T.
Others	99.94	0.06	0
North T.	16.67	**50**	**33.33**
South T.	1.73	**13.79**	**84.48**

Table 2: Five Topics Set, BN for a 90% of IG Confusion Matrix (5 utterances per block)

As we can see, there's a significant error source due to the confusion between the *"North Taxing"* topic and the *"South Taxing"* one. Air traffic control task at the Barajas airport presents few differences at functional level between *"North Taxing"* and *"South Taxing"* positions. The one and only distinction we can make from both control positions is the place of the airport where each one takes place. The immediate consequence is that both positions share most of their concepts. Thus the difference between both positions is reduced to the literal values which those concepts can take.

Exactly, for an 80% of IG both BN topologies share up to 7 concepts which mean more than 75% of all the input concepts for *"South"* case. This rises up to 13 for a 90% of IG which supposes almost 90% of shared concepts. So there are very few concepts that are specific of each topic. In this way we also have to take into account that surely those BN are sparsely trained in that case. For this reason we made the decision to merge both topics into one, the *"Taxing"* topic. Therefore, from now on we will consider just a four topic set.

6.2 BN models for different % of IG

In Figure 2 we present the overall topic identification results obtained for our new four topic set. Figure shows the identification error rate for three different IG percentages and for different block sizes. We have also added boundaries for a 95% confidence interval.

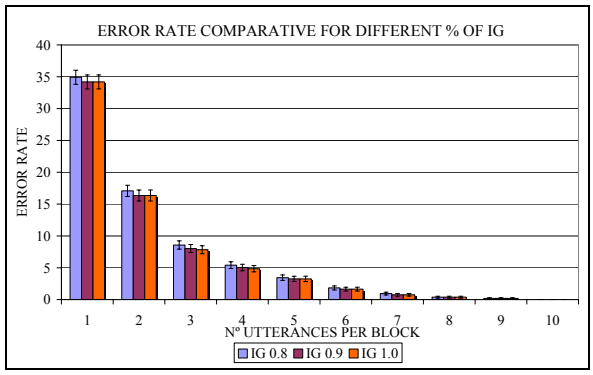

Figure 2: Overall topic identification error rate versus block size for different %'s of IG.

The obtained results match what we expected. The error rate is lower considering higher IG percentages and therefore employing more complex BN. Adding a greater number of concepts to the BN we can model more accurately the probabilistic dependencies between concepts and topics, at least if we have a suitable amount of data to train them. Since error boundaries throw no significant statistical difference between systems regarding the IG %, in the following sections we are going to assume the 80% of IG. This alternative minimizes the computational cost and shows similar performance.

Finally, the identification error rate also improves as we consider bigger block sizes since more evidences are available to identify the topic.

6.3 Detailed results for each topic

Next we are going to detail the results obtained for each topic for the case of BN models considering an 80% of IG (Figure 3).

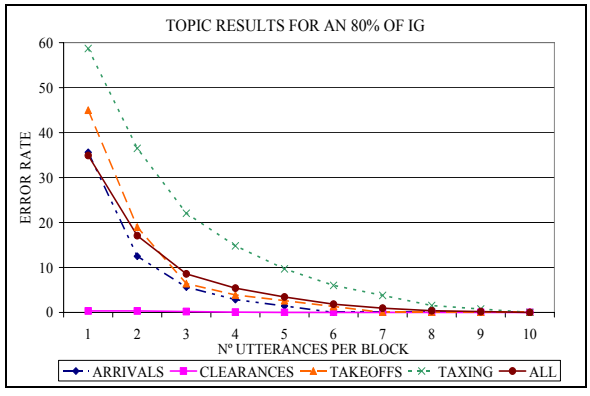

Figure 3: Detailed results for each topic for BN models for an 80% of IG.

From Figure 3, we can check an interesting effect which is that the error rate that we can observe for the *"Taxing"* topic is worse than for the rest of topics. We can explain these results using Table 3 where the average number of concepts per utterance for each one of the topics is presented.

	Average Nº of Concepts
Arrivals	4.44
Clearances	4.12
Takeoffs	4.14
Taxing	3.32

Table 3: Average number of concepts per utterance for each topic

It is possible to check that the average number of concepts is significantly smaller for the *"Taxing"* position. This has an important effect in terms of identification error rate since we have less available relevant information to carry out the topic identification. It is possible to notice that the effect of increasing the block size gives much better relative improvements in the case of *"Arrivals"* or *"Takeoffs"* positions than in the case of *"Taxing"*. For instance, if we increase the block size from one to two utterances we get a relative improvement of almost a 50% for *"Taxing"*, whereas for *"Arrivals"* this goes up to almost a 65%. This difference tends to be compensated as we use bigger block sizes.

Finally, we have also included the typical topic classification performance measures: *"recall"*, *R*, and *"precision"*, *P*. Regarding *"recall"* we consider the percentage of utterances (or blocks) correctly inferred by the BN for topic T_i out of all the true T_i utterances. Regarding *"precision"* we consider the percentage of utterances (or blocks) correctly inferred by the BN for topic T_i out of all the inferred T_i utterances (or blocks). Table 4 shows both results for a block size of five and also their combination (see Equation 3) as the *F-measure* (Rijsbergen, 1979) with β=1 in order to give recall and precision equal importance. We have assumed a block size of five as our reference because this is the smallest size for which the error rate is lower than 5%.

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

Equation 3: F–measure.

	Blocks	Recall	Precision	F
Arrivals	69	98.55	100.00	99.27
Clearances	1483	100.00	99.20	99.60
Takeoffs	76	97.37	96.10	96.73
Taxing	134	90.30	99.18	94.53
All	**1762**	**96.55**	**98.62**	**97.53**

Table 4: F-measure Results (block size 5)

7 Bayesian Classifier

Finally, in order to compare the results obtained using the BN, we decided to use a Bayesian classifier. We used a parametric method to estimate the class conditional probability density functions, pdf, from the training data. We used the same training set for this. We assume a Gaussian distribution for the pdf. The classes of our problem are the four topics we want to identify. We model each class with a Gaussian pdf of dimension four, (d=4). Thus, every parameter vector consists of four components that are the resultant perplexities, at phrase level, obtained by applying a specific 3-gram, or 1-gram, language model, LM, for each topic.

We have chosen such parameters because of their discriminative behaviour between classes and: i) perplexity is a normalized measure (independent of the phrase length), ii) 3-gram LM are suitable for such a phraseological task, iii) information used for both systems are different, as the LM uses information about the occurrence or not of each n-gram and, simultaneously, temporal information coded by the word-occurrence-order in the sentence (local history); meanwhile, the BN only takes into account the presence or the absence of an event.

The main disadvantages of this approach are: i) since LM works at word level, not at concept level, is necessary to train a different LM for each language and topic; this makes the LM more sensible to the data sparseness, ii) there is not information about the history of the phrases since we only take one isolated phrase at each time. Table 5 shows some statistics regarding the used LM.

	Spanish			
	No. Unigrams Train	No. Trigrams Train	Perplexity Test	OOV rate Test
Arrivals	115	726	4.50	0.97%
Clearances	760	11019	8.26	0.56%
Takeoffs	107	651	4.26	3.38%
Taxing	235	1535	10.98	1.64%
	English			
	No. Unigrams Train	No. Trigrams Train	Perplexity Test	OOV rate Test
Arrivals	62	230	4.20	4.52%
Clearances	272	2173	9.33	2.65%
Takeoffs	82	298	3.39	0.68%
Taxing	58	205	10.55	21.90%

Table 5: LM statistics for the Spanish and English database

Table 6 presents the confusion matrix for the results obtained using a 1-gram and 3-gram LM. The results for Spanish and English have been merged in each case. These results show that the Bayesian classifiers obtain an error rate of 33% for 1-grams and 19% for 3-grams.

We have included the 1-gram information, because, it is similar to the BN case in the sense that it does not take into account information about the word/concept order in the phrase. Besides, these values are interesting because they are similar to the cases when the BN use a block size of one or two (see Figure 3). So, whereas BN need 2 phrases to obtain a similar rate than the 3-gram LM, the last one only needs 1. This behaviour bears out the advantage of the LM, and it allows us to know intuitively the importance of include the temporal information into the BN.

	Unigrams			
	Arrivals	Clearances	Takeoffs	Taxing
Arrivals	**45**	22	6	0
Clearances	758	**487**	190	52
Takeoffs	11	28	**39**	2
Taxing	37	44	35	**22**
Global Error Rate	**33,35% ± 1.10**			
	Trigrams			
	Arrivals	Clearances	Takeoffs	Taxing
Arrivals	**63**	8	2	0
Clearances	124	**1240**	81	42
Takeoffs	2	11	**65**	2
Taxing	9	26	11	**72**
Global Error Rate	**19.01% ± 0.91**			

Table 6: Confusion matrix for the Bayesian classifier using 1-gram and 3-gram LM

8 *Conclusions and Future work*

In this work we have presented a BN approach for topic identification. We have obtained good classification results using a winner takes all evaluation scheme and a relatively small amount of data needed (less than 3.5% error rate with a block size of 5). Also, we made a comparison between the BN approach and a Bayesian classifier based on perplexity results using topic specific LM. These experiments showed the importance of the order of occurrence of the concepts in the phrase for this task. For this reason, our future work will be aimed to introduce this information, as well as train and evaluate new LM which include semantic and history information.

References

J.Ferreiros, J.Colás et al. "Controlling a HIFI with a continuous speech understanding system". The 5th Int. Conf. on Spoken Language Processing, 1998.

R.Pieraccini, E.Tzoukermann et al. "A Speech Understanding System Based on Statistical Representation of Semantics". Proc. ICASSP, 1992, pp. I-193 to I-196

D.Heckerman and E.Horvitz. "Inferring informational goals from free-text queries: A Bayesian Approach". Proc. 14th Conf. on Uncertainty in AI, 1998, pp. 230-238

H.M.Meng, C.Wai and R.Pieraccini. "The Use of Belief Networks for Mixed-Initiative Dialog Modelling", IEEE Transactions on Speech and Audio Processing, Vol.11, NO.6, pp. 757-773, 2003

H.M.Meng, W.Lam and C.Wai. "To believe is to understand". Proc. 6th Eur. Conf. Speech Communication and Technology, 1999

H.M.Meng, W.Lam and K.F.Low. "Learning Belief Networks for Language Understanding". Proc. of ASRU, 1999

INVOCA Project Synopses. Eurocontrol. Analysis of Research & Development in European Programs. http://www.eurocontrol.int

C.Huang and A.Darwiche. "Inference in Belief networks: a procedural guide". Int.Journal of Approximate Reasoning, 1994, pp.111-158

V.Rijsbergen, C.J. Information Retrieval. London. Butterworth, 1979