Statistical Decision Making applied to Text and Dialogue Corpora for Effective Plan Recognition

Manolis Maragoudakis, Aristomenis Thanopoulos and Nikos Fakotakis

Wire Communications Laboratory
Department of Electrical and Computer Engineering
26500 Rion, Patras, Greece
{mmarag,aristom,fakotaki}@wcl.ee.upatras.gr

Abstract. In this paper, we introduce an architecture designed to achieve effective plan recognition using Bayesian Networks which encode the semantic representation of the user's utterances. The structure of the networks is determined from dialogue corpora, thus eliminating the high cost process of hand-coding domain knowledge. The conditional probability distributions are learned during a training phase in which data are obtained by the same set of dialogue acts. Furthermore, we have incorporated a module that learns semantic similarities of words from raw text corpora and uses the extracted knowledge to resolve the issue of the unknown terms, thus enhancing plan recognition accuracy and improving the quality of the discourse. We present experimental results of an implementation of our platform for a weather information system and compare its performance against a similar, commercial one. Results depict significant improvement in the context of identifying the goals of the user. Moreover, we claim that our framework could straightforwardly be updated with new elements from the same domain or adapted to other domains as well.

1 Introduction

The majority of the dialogue systems that provide informational services such as news broadcasting, stock market briefing, route information, weather forecasting are system driven, meaning that the computer controls the process of interaction, expecting standardized, pre-defined queries from the user. By following this approach, the quality of the dialogue is deteriorated and circumscribed in narrow semantic limits, lacking of any mixed-initiative notion. In such systems, domain knowledge is handcrafted by an expert who should pay prominent attention during the design phase in order to create a representation that would be as robust as possible to the potential user's utterances variations. Such handcrafting of knowledge bases is infeasible for grappling with update or modification problems, since their structure is complex and domain specific. In addition, one should also consider that particularly for those interactions that occur via the telephony networks, direct and effective understanding of the intentions of a user is of great importance.

The term that has been introduced to describe the process of inferring intentions for actions from utterances is called "*plan recognition*" [3], [11]. Deriving the underlying aims can be assistive for a plethora of purposes such as predicting the agent's future behaviour, interpreting its past attitude creating a user model or narrowing the search space of a database query. Previous AI researchers have studied plan recognition for several types of tasks, such as discourse analysis [8], collaborative planning [10], adversarial planning [1], and story understanding [4].

For the present work, we propose a Bayesian network approach to modeling domain knowledge obtained from past dialogue acts and using it for interpreting the aims that lie beneath user's expressions. The structure as well as the conditional probability distributions of the networks are learned from manually annotated data derived from past dialogues. Furthermore, in order to effectively cope with unknown terms that may be found in a query, thus enhancing plan recognition accuracy, we estimate their semantic role from words similar to those comprising the system's vocabulary. The semantic similarities are obtained by applying a statistical algorithm, namely an information theoretic similarity measure [12] to raw text corpora.

We have applied the proposed method to a meteorological information system for the territory of Greece, which we call MeteoBayes and evaluated both its internal design issues and its performance against another, already operating system called METEONEWS[1] that can be accessed through the telephone. METEONEWS was designed and implemented using hand-coded domain knowledge and did not incorporate any plan recognition algorithm while in our system, MeteoBayes, we obtain domain knowledge from manually annotated dialogue parts and encode it into a group of Bayesian networks for the inference of a user's plan. Our experimental results depict significant improvement in the discourse quality, meaning the ability to quickly identify the user's aims. Moreover, we compare the complexity of each system's architecture and their potential ability either to be updated with new semantic elements or to be adapted to different domains.

2 Domain Description

Our Bayesian framework for plan recognition and dialogue managing for a weather information application, which from now on shall be referred as MeteoBayes, centers on conversations about goals typically handled by people located at the help desks of the weather information centers. We conducted an observational study of the weather forecast domain by recording 180 dialogue acts using the Wizard of Oz technique and by studying the log files of the METEONEWS telephone interactions. Through the reviewing process, we were able to identify a primal set of user goals as well as a key set of linguistic characteristics, relevant to the problem of detecting a user's demand. Observations revealed a group of 320 goals, with 48 of them mutually exclusive and exhaustive. One critical parameter that came into light during the reviewing process and needed to be taken into consideration is that users who interacted with the telephone service tended to clarify their goals from their initial utterances, while those who participated in the Wizard of Oz experiments in the laboratory were more

[1] Developed by Knowledge S.A. and Mediatel S.A.

abstract and haze in their plans. This phenomenon is caused from the telephone charging factor, which subconsciously forces the user to be more self-inclusive.

Upon completion of the domain analysis, five different semantic features were identified:

- Forecast: the concept of weather prognosis, including all possible variations such as general forecast, weather conditions in a specific area, wind bulletin, sea conditions, etc.
- Temperature: includes temperature and humidity report, heat or freeze alerts, etc.
- Time period: whether the weather forecast refers to today, tomorrow, from 3-6 days or for the whole week.
- Area: includes 10 big cities, 30 towns and 5 pelages in the Greek region.
- Land/Sea: whether the user is interested in a continental or thalassic area of a given place.

In addition to those features, we discovered considerable linguistic variability concerning the interactions. At times, users employed conventional phrases such as "*I would like to learn the temperature of Athens on Monday*" or "*Is tomorrow a sunny day in Rhodos?*". However, there was a significant number of abbreviated, more telegraphic queries such as "*Weather in Crete*" or "*Temperature today?*". Furthermore, there were cases where the goal was implied rather than stated. For example, the question "*Is the Rion-Patras canal accessible?*" implies the user's intention to be informed about the sea conditions in the thalassic area of the Gulf of the city of Patras.

3 Architecture

Our structure aims at the development of a dialogue system that would be independent of manual-coded domain knowledge and capable of easy adaptation to a different task. To achieve this, we have incorporated three separate modules, two off-line training systems and an on-line dialogue manager. The schematic representation that depicts their interconnection is shown in Fig. 1.

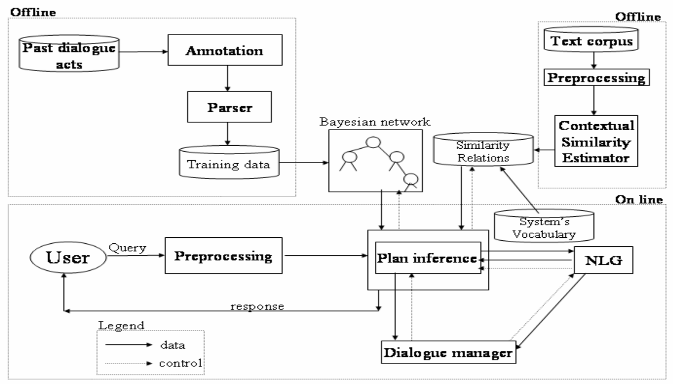

Fig. 1. Architecture of the proposed framework

3.1 Learning Domain Knowledge from Dialogue Acts

The off-line module that is responsible for the automatic acquisition of domain knowledge from the dialogue corpus operates as follows: Initially, we identify the primal set of semantic features that describe the task from the entire set of the past dialogue acts. This is actually the only phase where a domain expert is required. We have developed a parametric annotation tool in which such a specialist could dynamically define the input and output variables that enclose all the information which is suitable in order to perform an interaction. Regarding the annotation phase, note that not only nouns but linguistic elements that define a user's plan, such as temporal adverbs, present participles and adjectives, are annotated as well.

Upon completion of this procedure, a parser modifies the annotated dialogue corpus into a training set of lexical-semantic vectors that correspond to the mapping of the lexical parts of a user's utterance with the implied output semantic representation of his intentions. This training set will be used for the construction of the Bayesian networks that will encode the domain knowledge. These networks are learned using the following approach: Given a training set *D* that contains *n* different variables, the probability *P(B|D)* that a candidate network *B* is describing the data is estimated using the following equation [7]:

$$P(D \mid B) = \prod_{i=1}^{n} \prod_{j=1}^{qi} \frac{\Gamma(\frac{\Xi}{q_i})}{\Gamma(\frac{\Xi}{q_i} + N_{ij})} \prod_{k=1}^{ri} \frac{\Gamma(\frac{\Xi}{r_i q_i} + N_{ijk})}{\Gamma(\frac{\Xi}{r_i q_i})} \qquad \textbf{(1)}$$

where Γ is the gamma function, n equals to the number of variables and r_i denotes the number of values in *i:th* variable. q_i denotes the number of possible different data value combinations the parent variables can take, N_{ij} depicts the number of rows in data that have *j:th* data value combinations for parents of *i:th* variable, N_{ijk} corresponds to the number of rows that have *k:th* value for the *i:th* variable and which also have *j:th* data value combinations for the parents of *i:th* variable and Ξ is the equivalent sample size, a parameter that determines how readily we change our beliefs about the quantitative nature of dependencies when we see the data. In our study, Ξ equals to the average number of values variables have, divided by 2.

3.2 Learning Semantic Similarities of Words from Text Corpora

The purpose of the additional off-line module is to estimate the semantic role of words not appearing in our dialogue corpus from similar words. For this purpose, a database of semantic similarity relations is constructed from raw text corpora, on the basis of contextual similarity. Obtaining this information is very important in situations where an unseen term occurs in a user's phrase, permitting the flow of the interaction without having to ask the user for query reformulation.

Corpus-based automatic extraction of word similarities employs a central notion of contextual lexical semantics, namely that semantic properties of words can be defined "by their actual and potential context" [6] and therefore words that are found to share similar contexts along text corpora have certain semantic properties in common as well. Several measures for contextual similarity have been employed, such as the cosine [13], the minimal loss of mutual information [2], a weighted Jaccard similarity metric [9] and an information theoretic similarity measure [12]. [9] and [12] applied their measures on sets of syntactic dependencies from an analyzed corpus (i.e. <*word*$_1$, *syntactic_relation*, *word*$_2$>), which presupposes that a reliable syntactic parser will be available for the language of interest. Since our goal was to apply a generic and easily portable technique for the identification of similar concepts from raw text corpora, bypassing the need of sophisticated linguistic analysis tools, we considered a single contextual relation: Plain adjacency in text. Specifically, we considered as adjacent the words which are not separated by more than five content words nor by sentence boundaries. Therefore we achieve to correlate, besides semantically similar words, thematically relevant words, which is of considerable assistance in estimating the user's goal. We used Lin's metric [12], which assigns a similarity value between two words in the interval [0,1] and, considering a single contextual relation, is simplified to:

$$sim(w_1, w_2) = \frac{\sum_{(w)\in T(w_1)\cap T(w_2)} (I(w_1, w) + I(w_2, w))}{\sum_{(w)\in T(w_1)} I(w_1, w) + \sum_{(w)\in T(w_2)} I(w_2, w)} \qquad \textbf{(2)}$$

where T(w) is the set of words such that the mutual information of their co-occurrence with other words is positive[2], that is I(w,w′)>0, with

$$I(w, w') = \frac{c_{12} \cdot N}{c_1 \cdot c_2} \qquad \textbf{(3)}$$

where c_{12} is the frequency of the co-occurrence of w_1 and w_2, c_i the number of times w_i appears in the set of relations and N the number of the extracted relations from the corpus. From the obtained set of similarity relations we maintained only the N-best relations for every word of the system's vocabulary (we set N=100).

For this task we used the balanced ILSP/ELEYTHEROTYPIA corpus (1.6 million words), including news and articles obtained from a Greek daily newspaper.

The process so far is time-consuming but it is executed off-line and produces a set of relations of a rather small size. The on-line part is triggered by the occurrence of an out-of-vocabulary word, which is classified to the most plausible category using K-nearest neighbour classification (we set K=5).

In both learning processes the training text (i.e. the dialogue and newswire text corpora respectively) is pre-processed using a two-level morphological analyser that outputs the lemma of the word, in order to gather denser statistical data. Consequently, lemmatisation is also applied to the input query during the on-line process.

Additionally, we have included a date module that interprets any date format into the temporal periods we described in the domain description section. The purpose of the pre-processing stage is to identify parts of the input and match them to the domain lexicon items.

[2] We actually used I(w,w′)>1 in order to reduce the size of T(w) and thus computational cost and because values of I (mutual information) near zero indicate rather uncorrelated pairs.

3.3 The dialogue manager

The dialogue manager takes control after this stage and queries the appropriate Bayesian network in order to identify a plan. The response is guided to the plan inference module where it is interpreted and according to the degree of certainty about a user's plan, the NLG component replies either with the direct database answer or with verification and supplementary information sentences. In case an unknown word appears, the plan inference module consults the similarity relations database providing the system's vocabulary as a filter. The system's vocabulary term that mostly matches the unknown word (if any) is then considered to be the correct and the inference procedure is resumed.

Experimental Results

The evaluation of MeteoBayes focused on two different aspects. The former is the plan recognition performance with and without the semantic similarities knowledge base and the latter is to compare its architecture complexity with that of METEONEWS platform. Our approach was based on a set of 50 dialogue acts, provided by 10 users who were previously informed about the task and the possibility to imply their intentions than explicitly declaring them. This set was augmented by another set of 50 questions obtained by the log files of METEONEWS' past interactions. The total number of questions was 415. We separated this number into those questions where the goal was clearly defined and to those who was not. Let us denote the former set as Q_c, ($|Q_c|_{=295}$) and the latter set as Q_i, ($|Q_i|_{=120}$). As previously mentioned, the system was capable of identifying 48 mutually exclusive goals. Tables 1 and 2 tabulate the performance in terms of plan recognition accuracy without and with the semantic similarities module for both sets respectively. We manually set an empirical lower bound of certainty about the aim of a user to 60%. In case where the system did not meet this threshold, the plan inference module asked for a reformulated user question.

As can be observed in **Table 1**, 12 queries needed reformulation by the user, thus corresponding to 96% accuracy. From these queries, only 1 was unable to be understood even after reformulation, which corresponds to an error rate of 8%. On the contrary, in the Q_i set, the number of incorrectly identified queries is 41 (66%) and 7 items of this set could not be recognized even after the reformulation, resulting to an error rate almost 2 times bigger than that of the Q_c set. This performance is expected, since in the Q_i set, the intentions were implied by the users and not straightforwardly expressed. The reformulation stage did not necessarily involve the complete syntactic/semantic rephrasing of the question, but included spelling error checking as well.

The results obtained by incorporation of the semantic similarity database for unknown words (**Table 2**), indicate that there is actually little effect in the case that the user's goal is clearly defined while in the opposite case a significant improvement is accomplished. In the Q_i set particularly, the error rate after the reformulation stage drops by almost 45%.

Table 1. Plan recognition performance without the word similarities.

Category	*Amount*	*Accuracy*
Q_c	295	
Reformulation of a Q_c question	12	96%
Unidentified object	1	92%
Q_i	120	
Reformulation of a Q_i question	41	66%
Unidentified object	7	83%

Table 2. Plan recognition performance using the database of similarity relations.

Category	*Amount*	*Accuracy*
Q_c	295	
Reformulation of a Q_c question	13	95.5%
Unidentified object	0	100.0%
Q_i	120	
Reformulation of a Q_i question	21	82.5%
Unidentified object	2	90.5%

Concerning the evaluation of the architecture complexity, we examine the number and structure of the resource files needed, along with their flexibility to be updated since we cannot perform a straightforward comparison between the hardware and human-month effort required for both the METEONEWS and MeteoBayes development. From the METEONEWS point of view, there are 61 grammar files that interconnect in order to cover the weather forecast domain. These grammars are written in JavaScript grammar format and they utilize a template oriented approach. The parsing is performed by Philips Speech PERL 2000$^{\copyright}$ platform. As regards to MeteoBayes, the total number of lexical resource files is only 5, corresponding to the semantic features described in section 2. They contain the stem of the words that indicate each category. The average number of semantic relations we maintain for resolving unknown terms is 70 times the number of the system's vocabulary words. In case that new lexical elements should be included, the only step would be a simple addition in the corresponding lexical resource file while with METEONEWS, the same procedure would require the construction of a new grammar with potentially additional modifications to old ones, plus a new compilation of all.

In addition to the ability to be easily updated, we claim that the proposed framework can effortlessly be adapted to another domain. Once obtaining the dialogue acts and defining the semantic entities, the procedure of incorporating this knowledge into a dialogue system is uncomplicated. Only the Bayesian networks and the NLG responses will vary from task to task. The plan inference engine will remain the same.

Conclusion

The identification of a user's plan could contribute to the significant improvement of natural language human-computer interaction systems, since they enrich the dialogue quality, which is a very significant factor, particularly for telephone applications. Given the obvious high cost of manually encoding domain knowledge, this paper has presented a novel, Bayesian framework that aims to achieve plan inference ability without the need for hand-coded knowledge. More particular, we have introduced a platform that employs manually annotated past dialogue acts in order to obtain domain knowledge. This information is encoded into a group of Bayesian networks and is used for the user's goals identification procedure by a discourse manager module. Moreover, in order to cope with the complicated issue of unknown terms, an off-line system that learns semantic similarities from raw text was incorporated. The generated relations were used to replace the unknown word with a system's vocabulary term that had the most similar meaning. We have implemented the proposed approach by developing a weather information dialogue system, called MeteoBayes and compared it against another, hand-coded weather information system, named METEONEWS. Experimental results have depicted significant plan recognition accuracy. Moreover, the framework could straightforwardly be updated with new elements. Concluding, we argue that our method can be adapted to different domains with slight modifications.

References

1. Azarewicz, J., Fala, G., & Heithecker, C.: Template-based multi agent plan recognition for tactical situation assessment. In proceedings of the sixth Conference on Artificial Intelligence Applications (1989) 247–254
2. Brown P.F., DellaPietra V.J., DeSouza P.V., Lai J.C., Mercer R.L.: "Class-Based n-gram Models of Natural Language", Computational Linguistics, vol. 18 n°4, (1992). 467–479
3. Carberry L.: Incorporating default inferences into plan recognition. In Proc. 8th Nat. Conf. AI 1 (1990) 471–478
4. Charniak E., & Goldman, R. P.: A Bayesian model of plan recognition. Artificial Intelligence, 64 (1) (1993) 53–79
5. Clark H. H.: Using Language. Cambridge University Press (1996)
6. Cruse D.A.: Lexical Semantics. Cambridge University Press, Great Britain (1986)
7. Glymour C. and Cooper G. (eds): Computation, Causation & Discovery. AAAI Press/The MIT Press, Menlo Park (1999)
8. Grosz B. J. and Sidner C. L.: Plans for discourse. In: Cohen P. R., Morgan J. L., and Pollack M. E., (eds.): Intentions and Communication. Cambridge, MA: MIT Press. (1990) 417–444
9. Grefenstette G.: Explorations in Automatic Thesaurus Discovery, Kluwer Academic Publishers, Boston (1994)
10. Huber M. J., and Durfee E. H.: Observational uncertainty in plan recognition among interacting robots. In Working Notes: Workshop on Dynamically Interacting Robots, Chambery, France (1993) 68–75
11. Kautz H. A. & Allen J. F.: Generalized plan recognition. In Proceedings of AAAI (1986) 32–37
12. Lin D. 1998.: Automatic retrieval and clustering of similar words. In Proceedings of the COLING-ACL, Montreal, Canada
13. Schütze H.: Dimensions of Meaning, Supercomputing '92 (1992) 787–796