Domain knowledge acquisition and plan recognition by probabilistic reasoning

Manolis Maragoudakis, Aristomenis Thanopoulos, Kyriakos Sgarbas, Nikos Fakotakis
Speech and Language Technology Group
Department of Electrical and Computer Engineering
University of Patras
26500 Rion, Patras, GREECE
{mmarag,aristom,sgarbas,fakotaki}@wcl.ee.upatras.gr

Abstract. In this paper, a probabilistic framework for acquiring domain knowledge from heterogeneous corpora is introduced. The acquired information is used for intelligent human-computer interaction through the web. The application selected for the framework experimentation was education on issues of chemotherapy of nosocomial and community acquired pneumonia. The application was targeted on medical students as well as therapists and hospital personnel. Users could either learn about approximately 60 different antibiotics or ask for decision support on a given diagnosis using natural language for both cases. Contrasting to existing educational dialogue engines which use handcrafted knowledge of the application domain, our approach utilizes automatic encoding of the semantic model of the application, based on learning Bayesian networks from past user questions. The structure of the networks as well as the conditional probability distributions are computed automatically from dialogue corpora, thus eliminating the tedious process of manual insertion of domain knowledge. Furthermore, we attempt to overcome the significant issue of limited vocabulary by incorporating a methodology which estimates semantic similarities of words not found within the system's vocabulary and probabilistically associates them with those who appear. The evaluation of the platform was performed against an existing medical educational system called DIKTIS, the architecture of which is based on manually embedded domain knowledge. Obtained results depict significant improvement in the context of effectively identifying the underlying goals of a user. The presented approach demonstrates a 24% recognition improvement using the automatic domain knowledge extraction engine, augmented with the unknown terms resolving module. Furthermore, it is stressed that the proposed framework can straightforwardly be updated with new lexical or semantic elements either manually inserted or automatically obtained from texts.

1. Introduction

A major aim of medical informatics is to improve the quality of the interaction between therapists and medical decision support dialogue systems. The majority of existing dialogue systems that provide informational services or decision support 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 interaction is deteriorated and circumscribed in narrow semantic limits, lacking of any mixed-initiative notion. In such systems, domain knowledge is typically handcrafted by an expert who should pay prominent attention during the design phase in order to structure a semantic representation that would be as robust as possible to the potential user's question variations. Such handcrafting of knowledge bases is infeasible for grappling with the linguistic (both lexical and semantic) diversity of a domain.

The term that has been introduced to describe the process of inferring intentions for actions from questions is called "*plan recognition*" [3,12]. 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, speeding the discourse process up or narrowing the search space of a database query. Previous AI researches have studied plan recognition for several types of tasks, such as discourse analysis [10], collaborative planning [11], adversarial planning [1], and story understanding [4]. Plan recognition is considered to be the most significant step in the process of natural language understanding [17]. Plans are synthesized by a rational agent with some beliefs, preferences, and capabilities, that is, a mental state. Knowing the agent's mental state and its rationality properties strongly constrains the possible plans it will construct. (The degree of constraint depends on the power of the rationality theory we adopt.) Plan recognition ability is strongly related to the richness of domain knowledge representation. The automatic generation of such knowledge from data is one of the main aims of our approach. For the present work, we propose a Bayesian method for modeling

domain knowledge from dialogue data. Our framework is Bayesian in that we start from a causal theory of how the agent's mental state causes its plan, executing its plan causes activity and we reason from observed effects to underlying causes. Our recognizer has uncertain a priori knowledge about the agent's mental state, the world state, and the world's dynamics, which can be summarized (at least in principle) by a probability distribution. It then makes partial observations about the world, and uses this evidence to induce properties of the agent and its plan. Dialogue data are a prolific source for obtaining the above parameters needed for effective understanding since they incorporate the mental states of many users, thus encoding un-biased semantic knowledge. We provide Bayesian analysis regarding the influence each lexical item poses to each other, by considering either the naïve Bayes or the Bayesian networks approach. However, due to the fact that new queries that contain unknown terms, not found within the training dialogue corpus, may appear, we apply a novel approach that estimates semantic similarities of words from generic and domain specific raw corpora. The statistical extraction of this information and its incorporation to the domain knowledge modulates the other main goal of our approach.

2. Domain specification and the DIKTIS project

The term pneumonia, either nosocomial or community acquired [9], refers to the inflammation of the lung parenchyma, due to bacterial, viral, fungal, parasitic causative micro organisms, that is accompanied with clinical and radiological features of one or more lung densities. The term “nosocomial pneumonia” corresponds to those types of pneumonia which develop 48-72 hours upon hospital admittance. Patients which are involved in nosocomial pneumonia are categorized according to their pathophysiologic state [8]. Community acquired pneumonia is formally defined as an acute infection of the pulmonary parenchyma that is associated with at least some symptoms of acute infection. It occurs in a patient who is not hospitalized or residing in a long-term-care facility for more than 14 days before the onset of symptoms. The DIKTIS [15] project aimed at developing a medical expert system that could provide assistance to the specialist’s decision, for the task of treating pneumonia. It was funded by the Hellenic general secretariat for research and technology. Since September 2001, it has been installed in the “Evangelismos” hospital in Athens, Greece where it remains operational. The specialists interact with DIKTIS in Modern Greek, by using written natural language. The domain knowledge was manually inserted by medical experts. We augmented its functionality allowing users to interact through the Web, using the understanding engine which builds semantic knowledge from available medical data. The resulting module allows medical student and experts to connect to it from any point.

3. Benefits of automatic domain knowledge acquisition

Throughout the process of knowledge representation from domain experts, they should pay prominent attention in order to cover a broad variety and diversity of the semantic categories of the task at design time, thus danger of not being able to fully represent the aspects of the domain may lurk. Furthermore, since the final goal is to correctly identify the expressions of a user, stated in natural language, one should also take the wide range of linguistic aberrations into account. Even in the ideal case that the extracted knowledge base fully covers the linguistic and semantic topics, since dialogue systems interact with new users, the probability of newly formed utterances to appear is augmented.

The proposed approach determines the underlying semantic knowledge from data, particularly from past dialogue acts. We are motivated by the fact that taking into account many users, results in a more general perspective of the application domain (since a greater, non-biased diversity of expressions appears). The resulting framework is updatable, meaning that recently interaction histories could be used in order to adjust the structure of the semantic knowledge or even enrich it with new lexical elements. However, using dialogue data as a source for inferring domain knowledge, poses a significant issue; how to deal with words that were absent from the dialogue corpus, thus haven’t been considered by the system’s vocabulary. The most apparent methodology is to collect a very large set of dialogues. This approach is impractical since collecting and more specifically annotating such a set is considered to be of significant difficulty. For the present work, we endeavour to overcome this problem by introducing a statistical method for extracting information about the similarity of words from text corpora. The extracted information is incorporated to the Bayesian domain knowledge

acquisition framework, thus assisting in reducing the cases where unknown terms cause the need for query reformulations.

4. Probabilistic framework of semantic representation

Semantic interpretation of an input query could be considered as the process of searching for the optimal (most probable) semantic interpretation through the space of candidate similar semantic interpretations for a specific domain. The set of provided lexical items - usually named as "keywords" - define the meaning of a query. In the more general case, one would claim that the above mentioned hypothesis space actually contains all the semantic categories of the domain. However, during search process, those who do not resemble the candidate semantic interpretations are superseded. In our approach, a probabilistic model for modeling semantic disambiguation is defined over a search space *H*T*, where *H* denotes the set of possible lexical contexts that could be identified within an input query $\{h_1,\ldots,h_k\}$ or "input variables" and *T* denotes the set of the allowable semantic interpretations of that question $\{t_1,\ldots,t_n\}$. Using Bayes' rule, the probability of the optimal interpretation T_{opt} equals to:

$$T_{opt} = \underset{T\hat{I}(t_1,\ldots,t_n)}{\operatorname{argmax}}\, p(T|H) = \underset{T\hat{I}(t_1,\ldots,t_n)}{\operatorname{argmax}}\, \frac{p(H|T)p(T)}{p(H)} = \underset{T\hat{I}(t_1,\ldots,t_n)}{\operatorname{argmax}}\, p(H|T)p(T) \quad (1)$$

Approximations of the probability distributions of equation (1) deal with the trade-off between computational complexity and efficiency [12]. For a given observation sequence of input variables $\{h_1,\ldots,h_k\}$, equation (1) is modified into:

$$T_{opt} = \underset{T\hat{I}(t_1,\ldots,t_n)}{\operatorname{argmax}} = p(t_i)p(h_1,\ldots,h_k|t_i) \quad (2)$$

There are two possible assumptions that can be considered from this point, regarding how do the lexical input items are considered to be; either to be regarded as independent of each other or to take into account that there is some specific kind of dependency among all variables or a subset of them. The first assumption denotes that given an utterance that contains lexical elements which trigger a semantic representation, those elements are not influencing one another. Returning to equation (2), if we assume that each lexical item is independent of each other, we adopt the naïve Bayes approach, while in the case we take the dependency of lexical items into consideration, we apply the Bayesian networks approach. Following former one, we consider elements $h_1,\ldots,h_k$ as conditionally independent, so equation (2) becomes:

$$T_{opt} = \underset{T\hat{I}(t_1,\ldots,t_n)}{\operatorname{argmax}}\, p(t_j)\prod_{i=1}^{k} p(h_i|t_j) \quad (3)$$

Bayesian networks are capable of effectively coping with the over-simplifying naïve Bayes restriction, since they allow stating conditional independence assumptions that apply to all or to subsets of the variables. They are characterized as a significant knowledge representation and reasoning tool, under conditions of uncertainty. We denote a network *B* as a pair *B*=<*S*,*P*> [16] where *S* is a DAG whose nodes correspond to the variables of *H*. *P* refers to the set of probability distributions that quantifies the network. The unique joint probability distribution over *H* that a network *B* describes can be computed using equation (4). The optimal value T_{opt} of a class variable (semantic interpretation) equals to equation (5).

$$p_B(H_1,\ldots H_n) = \prod_{i=1}^{n} p(H_i|\text{parents}(H_i)) \quad (4)$$

$$T_{opt} = \underset{T\hat{I}(t_1,\ldots,t_n)}{\operatorname{argmax}}\, p(t_j)\prod_{i=1}^{k} p(h_i|\text{parents}(h_i),t_j) \quad (5)$$

We have previously argued that automatically extracted knowledge is more beneficial than manual, regarding the natural language understanding task. For that reason, we consider learning the structure and the parameters of a network from available data as a more appropriate methodology. In order to infer on the most probable network structure from data, we followed the [5] approach. The dialogue acts data used in this approach were collected from the past dialogue acts of DIKTIS system at the "Evangelismos" hospital and at the school of Medicine of the University of Patras. More specifically, 300 DIKTIS past interactions with the hospital doctors along with 200 utterances provided by 20 different users formed the dialogue data set. The users were senior students at the school of Medicine, previously informed on the task of interacting with DIKTIS using natural language. The data set was manually annotated.

5. Estimating the Semantic Role of Unknown Words from Raw Text Corpora

An apparent problem in dialogue-corpus-based approaches for query-understanding is that the system's vocabulary is restricted by the limited dialogue corpus used for training. In order to acquire more lexical knowledge the semantic role of unknown words is estimated by their similar words, looking into a database of lexico-semantic similarities, acquired via the analysis of the contextual behavior of words in larger collections of texts on the domain of interest. For this purpose we used 3 Greek corpora: a) A collection of the medicine descriptions of Hellenic National Organization for Medicines (HNOM), (426.000 words). b) The ILSP/Eleftherotypia newswire corpus (3 million words). c) A corpus automatically extracted from WWW by an intelligent agent (4.2 million words). In order to extract lexical similarities three different characteristics of text were exploited:

1. Words with similar lexico-semantic properties tend to appear in similar lexical environments; inversely, similar contextual environments suggest possible lexico-semantic similarity. In order to measure contextual similarity we employed the information-theoretic similarity metric (sim_{LIN}) proposed in [13]. We took into consideration two types of contextual relations: Immediate precedence and succession in local context, allowing only interpolation of functional words.
2. Certain text passages in domain-focused corpora discuss a specific subject and they contain mostly words related to the subject. In the HNOM corpus paragraphs related to the medicine descriptions are labeled according to the respective topic (e.g. indications, side-effects, etc.) We identified statistical correlation between topics and words with likelihood ratio [7].
3. Lists and conjunctive expressions typically include similar elements, as in the sentence: “The most common side-effects are headache, dizziness and dyspepsia”. Therefore, we expect that words that recurrently co-occur in such expressions are probably semantically similar. In order to identify words which co-occur in such expressions much more often than expected by chance we employed likelihood ratio as well.

User’s spelling errors also bring in unknown words, complicating system understanding. A similar problem is emphasized by the inability of the morphological analyzer to process words absent from its lexicon, as a query word will not be matched against a corpus word of the same lemma, albeit morphologically similar. We employed a morphological similarity estimator to confront with both problems, using again the information-theoretic similarity metric ($sim_{LIN\text{-}morph}$) proposed in [13]. When an unknown word occurs in the user’s query, the system tries to find it in the ranked word similarity lists obtained by contextual similarity and co-occurrence in conjunctive expressions. If no known similar word is found, it searches for the more associated topic. If no topic is found, it calculates the string similarities with the domain vocabulary words to find the morphologically most similar word, provided their similarity exceeds the threshold $sim_{LIN\text{-}mprph} = 0.9$.

6. Experimental results

For the evaluation of the proposed method, we conducted two different experiments. The purpose of them was to measure the performance of semantic interpretation recognition with and without the semantic similarities knowledge base, using past dialogue acts as well as examples from a runtime operation of the system, in which a set of 10 users provided 10 questions each. Regarding the former case, the set of 500 annotated dialogue examples was partitioned into equal parts of 100 queries. The 90% of those parts was used for training material while the remaining 10% comprised the testing set. This process was repeated 10 times, using different parts of the query set for training and evaluation each time. We used both the naïve Bayes and Bayesian network inference algorithms. The performance was measured by the average identification accuracy over the 10 repetitions. The accuracy was calculated as the number of correctly identified semantic categories by the number of questions in the open test set. As a comparison measure, the recognition accuracy of the original rule-based DIKTIS’ configuration is also provided. The summary of the best experimental results is tabulated in Table 1, while, Fig. 1, depicts the progress of error rate using different sizes of training material. We observed that both naïve Bayes and Bayesian network modeling of semantic knowledge acquisition outperform the rule-based original approach of DIKTIS, introducing an advantage ratio of 7%-12%. Furthermore, a significant improvement in terms of error rate is observed when we incorporated the algorithm for searching over the semantic similarities databases.

Table 1: Ten fold cross validation % accuracy of the understanding module for the original DIKTIS configuration and all proposed methodologies plus the corresponding advantage ration against the original DIKTIS' performance.

	Original DIKTIS	*Naïve Bayes*	*Bayesian network*	*Naïve Bayes and similarities*	*Bayesian networks and similarities*
Accuracy%	*52,70±0,00*	*56,39±2,78*	*62,19±4,21*	*58,50±3,04*	*65,35±4,56*
Adv. ratio	*-*	*0,07*	*0,18*	*0,12*	*0,24*

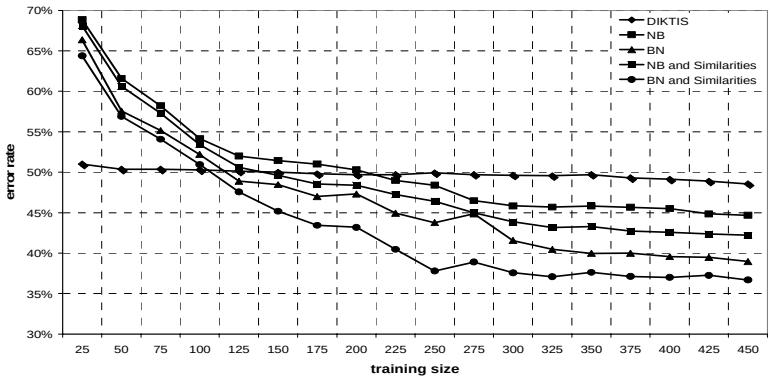

Fig. 1. Percentage of error rate in the test set 10% of the training size for fluctuant training text.

An increase is also introduced utilizing the naïve Bayes with the similarities databases. An interesting observation is that Bayesian networks appeared to behave better than naive Bayes, which justifies our previous assumption on the dependency relationships among lexical features of a sentence.

Table 2: Query understanding performance of the proposed framework using Bayesian network modeling with and without semantic similarities databases.

Category	**Questions**	**Error rate**
Without similarities	100	
Reformulation of a Q_c question	36	36% (39/100)
Unidentified object	9	25% (9/36)
Using similarities	100	
Reformulation of a Q_i question	28	28% (28/100)
Unidentified object	4	14%(4/28)

Regarding the latter case, our aim was to measure the performance of our system in the real world, thus we conducted a qualitative evaluation [2] by introducing the system to 10 medical students, different than those who participated in the dialogue data collection phase, which were asked to provide 10 questions each. In order to estimate performance of the system, we measured the number of correctly identified questions as well as how many questions of the 100 set needed reformulation. Analysis of this task (Table 2), acquiesced to the claim that incorporating semantic similarities knowledge results in recognition improvement.

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 and Applications, (1989) 247-254.

[2] Beach S., Gevarter W. Standards for evaluating expert system tools, Expert Syst Appl 2 (1991) 259-267.

[3] Carberry L. Incorporating default inferences into plan recognition, in Proceedings of the 8th Nat. Conf. AI, (1990) volume 1, 471-478.

[4] Charniak E., & Goldman, R. P. A Bayesian model of plan recognition. *Artificial Intelligence*, 64 (1993) 53-79.

[5] Cooper J., Herskovits E. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9 (1992) 309-347.

[6] Dermatas E., Kokkinakis G. Automatic stochastic tagging of natural language texts, *Computational Linguistics*, 21(2) (1995) 137-163.

[7] Dunning T. Accurate methods for the statistics of surprise and coincidence, *Computational Linguistics* 19.1 (1993)

[8] Fagon J. Chaste J. Hance A. Nosocomial pneumonia in ventilated patients: a cohort study evaluating attributable mortality and hospital stay. *Am J Med*, 94 (1993) 281-288.

[9] Farr BM. Prognosis and decisions in pneumonia, *N Engl J Med*, 337 (1997) 288-290.

[10] 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.

[11] Huber M. J., and Durfee E. H. Observational uncertainty in plan recognition among interacting robots, in working notes of the Workshop on Dynamically Interacting Robots, (Chambery, France, 1993) 68-75.

[12] Kautz H. A. & Allen J. F. Generalized plan recognition, in Proceedings of AAAI-86 (1986) 32-37.

[13] Lin D. Automatic retrieval and clustering of similar words, in: Proceedings of the COLING-ACL'98 (Montreal, Canada, 1998) 768-774.

[14] Manning C., Schütze H. Foundations of Statistical Natural Language Processing (MIT Press, Cambridge, 1999).

[15] Maragoudakis M., Kladis B., Tsopanoglou A., Sgarbas K., Fakotakis N. Natural language in dialogue systems. A case study on a medical application, in Proceedings of the International Conference in Human–Computer Interaction, PCHCI, Patras Greece, (2001) 197-201.

[16] Pearl J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (San Mateo, CA: Morgan Kaufmann, 1988).

[17] Zukerman I., Litman D. Natural language processing and user modeling: synergies and limitations, *User modeling and user adapted interaction*, 11 (2001) 129-158.