A Knowledge-Based Approach for Automatic Generation of Summaries of Behavior

Martin Molina and Victor Flores

Department of Artificial Intelligence, Universidad Politécnica de Madrid,
Campus de Montegancedo S/N 28660 Boadilla del Monte, Madrid, Spain.
{mmolina,vflores}@fi.upm.es

Abstract. Effective automatic summarization usually requires simulating human reasoning such as abstraction or relevance reasoning. In this paper we describe a solution for this type of reasoning in the particular case of surveillance of the behavior of a dynamic system using sensor data. The paper first presents the approach describing the required type of knowledge with a possible representation. This includes knowledge about the system structure, behavior, interpretation and saliency. Then, the paper shows the inference algorithm to produce a summarization tree based on the exploitation of the physical characteristics of the system. The paper illustrates how the method is used in the context of automatic generation of summaries of behavior in an application for basin surveillance in the presence of river floods.

1 Introduction

General techniques for automatic summarization usually simulate human reasoning such as abstraction or relevance reasoning. For example, techniques for event summarization include exploiting the saliency of events (with domain properties or statistics), abstracting events from collections of events, and integrating events based on semantic relations [1]. A particular application of automatic summarization is report generation in the context of control centers where the behavior of a dynamic system is supervised by human operators. Here, operators make decisions on real-time about control actions to be done in order to keep the system behavior within certain desired limits according to a general management strategy. Examples of these dynamic systems are: a road traffic network, the refrigeration system of a nuclear plant, a river basin, etc.

In this context, physical properties of dynamic systems provide specific criteria to formulate more specific techniques for summarizing and relevance reasoning. According to this, we present in this paper a knowledge-based approach that can be used to generate summaries in the context of surveillance of the behavior of dynamic systems. In the paper we analyze the type of knowledge and representation required for this type of task and we describe the main steps of an inference algorithm. We illustrate this proposal with the case of a particular application in the field of hydrology where thousands of values are summarized in single relevant states. At the end of the paper we make a comparative discussion with similar approaches.

In "Artificial Intelligence: Methodology, Systems, and Applications", J. Euzenat and J. Domingue (Eds.). 12th International Conference, AIMSA 2006, LNAI 4183, pp. 265 – 274. Varna, Bulgaria, September 2006.

2 The method for summarization

In automatic summarization two separated tasks can be considered: (1) *summarize* the most important information (i.e., *what* to inform) and (2) *present* the information using an adequate communication media according to the type of end-user (*how* to present the information). This paper describes our approach for the summarization task and, then, the paper illustrates how it is related to the presentation task in a hydrologic domain.

According to modern knowledge engineering methodologies [2], we have designed a method conceived with a set of general inference steps that use domain specific knowledge. In the following, we first describe the types of domain knowledge used in the method: (1) *system model*, (2) *interpretation model* and (3) *salience model*. Then, we describe the general inference as an algorithm that uses these models with a particular control regime.

2.1 The system model

The *system model* is a representation of an abstraction about behavior and structure of the dynamic system. Our method was designed to simulate professional human operators in control centers with partial and approximated knowledge about the dynamic system. Therefore, the system model was conceived to be formulated with a qualitative approach instead of a precise mathematical representation with quantitative parameters.

The representation for the system model is a particular adaptation of representations and ontologies used in qualitative physics (e.g., [3] [4] [5] [6]). In the model, a detailed hierarchical representation of the structure is followed to support summarization procedures. However, since the system model is not used for simulation of the dynamic system, the behavior is represented with a simpler approach.

In particular, the structure of the dynamic system is represented with a set of *components* $C = \{C_i\}$. Each component represents a physical object of the system such as a reservoir, river or rainfall area in the basin. In a given moment, a component C_i presents a qualitative *state*. Each component C_i is also characterized in more detail with quantitative measures corresponding to physical *quantities* $Q_1, \ldots, Q_k$ (e.g., water-level and volume of a reservoir). Components are related to other components with the relations *is-a* and *member* (user-defined relations can be also used to consider domain-specific relations). A *parameter* is a tuple $P_i = <C_i, Q_i, F_i, T_i>$ that represents a physical variable defined by the component C_i, the quantity Q_i, optionally a *function* F_i (e.g., as average time value, time derivative, maximum value, etc.) and optionally a *temporal reference* T_i (temporal references are *time points* or *time intervals*). An example of parameter is `<Casasola, level, max, [18:00, 21:00]>` which means the maximum value of the water level in the Casasola reservoir between 18:00 and 21:00 hours.

The model includes also a simplified view of the system behavior represented with *causal relations* between physical quantities. These relations can include labels such as temporal references about delay or type of influence (M^+ or M^-, i.e. increasing or decreasing monotonic functions, etc.). *Historical values* also help to represent

information about behavior (e.g., average values, maximum historical values, etc.). Figure 1 shows a simplified example in the hydrologic domain that summarizes this representation.

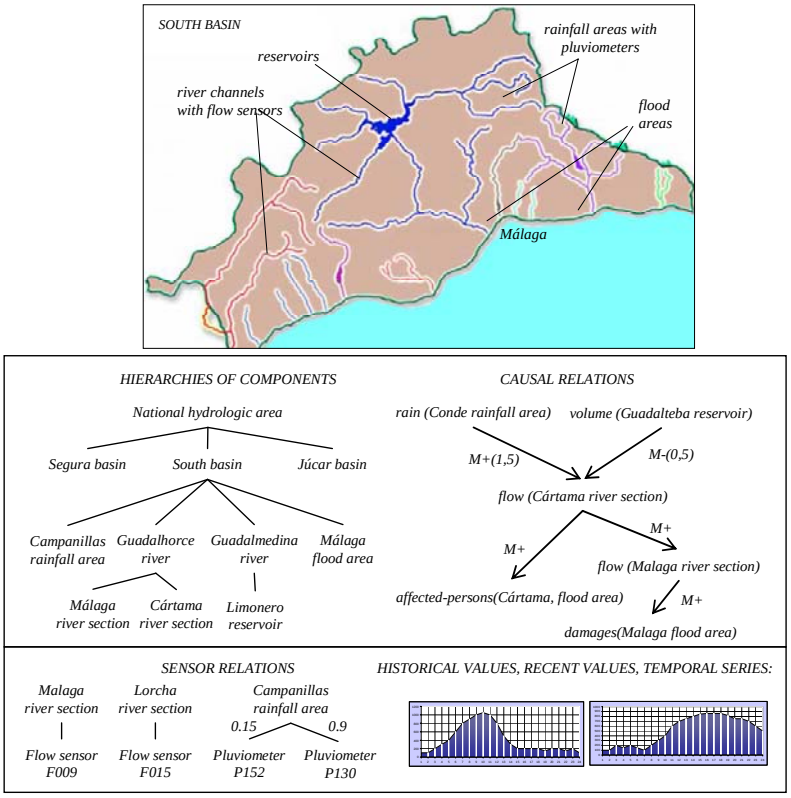

Figure 1: Summary of the representation for the dynamic system in the domain of hydrology.

2.2 The interpretation model

The interpretation model expresses how to determine the qualitative state of every node in the hierarchy of components. For the case of single components their state is determined directly by conditions about parameters. Normally, this is formulated with conditions about limit points that define the quantitative space corresponding to the state. This can be formulated by *qualitative interpretation rules*, i.e. sentences with the following format (x is a component and y_j are parameters):

$$\forall x, y_1, \ldots, y_n (type(x,a) \wedge param(x,y_1) \wedge \ldots \wedge param(x,y_n) \wedge COND_k(y_1, \ldots, y_n) \rightarrow state(x,b))$$

where *type(x, a)* means that the type of component x is a, *param(x,y)* means that y is a parameter of x, $COND_k$ is a logical expression about the values of parameters y_1, …, y_n and *state*(x, b) means that the state of the component x is b. An example in natural language is: *the state of a reservoir is near-limit-increasing if its volume is between 90% and 100% of its capacity and the time derivative of the volume is positive*.

For the case of complex components their state is determined by conditions about the state of simpler components. This can be formulated by *aggregation rules*, i.e. sentences based on the following format (*x, y* are components):

$$\forall x, y \ (type(x,a) \wedge member(y,x) \wedge type(y,b) \wedge state(y,c) \rightarrow state(x,d))$$

where *member(y,x)* means that the component *y* is member of the component *x*. With this type of rules, a particular component could deduce different states based on the states of its different members. So these sentences must be interpreted following a particular control mechanism based on relevance as it is described in the following section. An example in natural language is: *the state of the basin is damages if there is a flood-area of the basin that presents the state of agricultural-losses.*

The interpretation model also is used to formulate how the value of a parameter *x* is computed based on the values of other parameters y_1, y_2, …, y_n (when *x* is not directly measured by sensors). This can be expressed with functional sentences where each sentence associates to a parameter *x* a function applied to the other parameters y_1, y_2, …, y_n. The function is taken from a library that includes arithmetic functions, statistical functions for both temporal and component abstraction, etc. An example of this sentence in natural language is: *the storage percent of a reservoir is the current volume of the reservoir multiplied by 100 and divided into the capacity of the reservoir.*

2.3 The salience model

The salience model represents a kind of control knowledge to determine when certain event is relevant to be reported to the operator. In general, we consider a relevant event as a significant deviation of the desired state established by the goals of the management strategy of the dynamic system. This definition is valid to report the relevant information about the behavior of the dynamic system during a long period of time. However, when operators monitor on real time the behavior of the system, we consider the notion of relevance as follows:

> **Definition.** A *relevant event* is an event that (1) changes with respect to the immediate past and (2) produces a change (now or in the near future) in the distance between the state of the dynamic system and the desired state established by the management goals.

The implication of this definition is that, in order to evaluate the relevance of facts, it is necessary to predict the final effect of state transitions. However, based on our assumption for system modeling, we follow here a simplified and efficient approach with approximated knowledge for the system behavior. According to this, the representation of relevance establishes when a state can affect to the management goals, using a heuristic approach that summarizes sets of behaviors. This is formulated as logic implications that include (1) in the antecedent, circumstantial conditions about states of components and (2) in the consequent, the state of a component that should be considered relevant under such conditions. The general format is (*x* and y_j are components):

$$\forall x, y_1, \ldots, y_n (type(x,a) \wedge REL_k(x, y_1, \ldots, y_n) \wedge state(y_1, b_1) \wedge \ldots \wedge state(y_n, b_n) \rightarrow relevant(state(x,c))$$

where $REL_k(x,y_1,\ldots,y_n)$ relates a component x with other components $y_1, \ldots, y_n$, according to physical properties (for instance a relation that represents the reservoirs that belong to a river). Thus, in hydrology, light rain is normally considered non relevant except, for example, if the weather forecast predicts heavy rain and the volume in a reservoir downstream is near the capacity.

Our notion of relevance gives also criteria to establish order among relevant events. This can be done with sentences that represent heuristic knowledge defining priority between two states based on their impact on the management goals. The representation uses conditional sentences that conclude about preference between states (represented by $A > B$, A is more relevant than B) with the following format (x and y are components):

$$\forall x, y(type(x,a) \wedge type(y,b) \wedge COND_k(x,y) \rightarrow state(x,a) > state(y,b))$$

where $COND_k$ is a logical expression (possibly empty) about the components x and y. For example, in hydrology, this allows to establish that heavy-rain at certain location x_1 is more relevant than the same rain at location x_2. It also allows formulating a general priority scheme like: *damages* > *volume* > *flow* > *rain* > *weather-forecast*.

It is important to note that this priority scheme plays the role of control knowledge in the complete model. The aggregation rules of the interpretation model are used to determine the state of components based on the state of simpler ones. However, to avoid contradictory conclusions, these sentences need to be applied according to certain control mechanism. The relevance priority is used here for this purpose taking into account that sentences that interpret qualitative states with higher priority are applied first.

2.4 The general inference

The general inference exploits the physical system properties (e.g., causal relations, member relations and changes in qualitative states) together with domain knowledge about relevance to produce the summary. In particular it performs a linear sequence of the following inference steps:

1. *Interpret*. For every single component its qualitative state is computed using as input the qualitative interpretation rules and the measures of sensors.
2. *Select*. Relevant states are selected. For every single component, the relevance of its state is determined by using the saliency model according to the following definition. A state $S(t_i)$ in the present time t_i of a component C is relevant if (1) the state $S(t_{i-1})$ of component C in the immediate past changes, i.e., $S(t_i) \neq S(t_{i-1})$, and (2) the predicate $state(C, S(t_i))$ is deduced as relevant according to the domain-dependent rules of the salience model. Let $R = \{S_1, S_2, \ldots, S_n\}$ be the set of relevant states.
3. *Sort*. The set R of relevant states is sorted according to the domain-based heuristics of the salience model. More relevant states are located first in R.
4. *Filter*. Less relevant states that correspond to the same physical phenomenon are removed. For each state S_i in R (following the priority order in R) a second state S_k is removed from R if (1) S_k is less relevant than S_i (i.e., S_k is located after S_i in

R), and (2) S_k is member of *causes*(S_i) or S_k is member of *effects*(S_i). Here, *causes(X)* and *effects(X)* are the sets that respectively contain all the (direct or indirect) causes and effects of *X* based on the causal relations of the system model.

5. *Condense*. The states of similar components are condensed by (1) *aggregation* and (2) *abstraction*. States of components with the same type are *aggregated* by the state of a more global component by using the aggregation rules of the interpretation model. Here, the salience model is used as control knowledge to select among candidate rules as it was described in the previous section. In addition to that, states of components of different type are *abstracted* by the most relevant state using the priority order in *R*. This produces what we call a *summarization tree*.

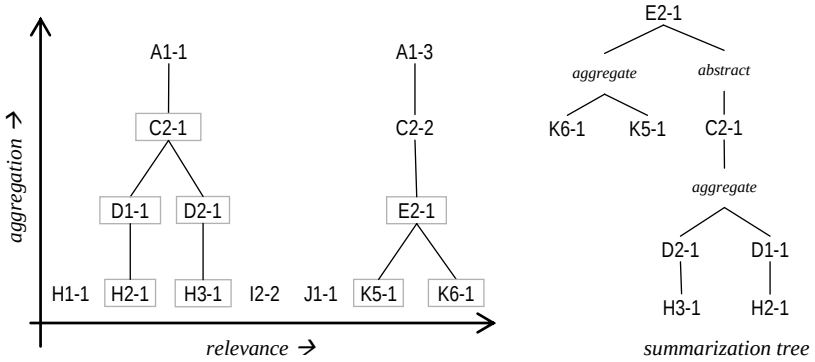

Figure 2: Example of summarization tree.

The example of figure 2 shows a summarization tree corresponding to a set of relevant states. In the example, the graphic at left hand side shows a partial search space. At the bottom, there are states of single components (K6-1 means the state 1 of component 6 of type K). The horizontal axis shows the relevance order (e.g, K6-1 is more relevant than K5-1). The squared states correspond to the elements of *R* = {K6-1, K5-1, H3-1, H2-1}. Upper nodes in these hierarchies are potential states inferred by aggregation rules. The corresponding summarization tree is presented at the right hand side. In this tree the most relevant and aggregated state is represented by the root E2-1.

3 Application in Hydrology

The previous general approach has been applied to the field of hydrology. In Spain, the SAIH National Programme (Spanish acronym for Automatic System Information in Hydrology) was initiated with the goal of installing sensor devices and telecommunications networks in the main river basins to get on real time in a control center the information about the hydrologic state. One of the main goals of this type of control centers is to help to react in the presence emergency situations as a consequence of river floods. The management goals in this case are oriented to

operate reservoirs to avoid problems produced by floods and, if problems cannot be avoided, to send information to the public institutions in order to plan defensive actions. Here, the generation of summaries of behavior is oriented to report relevant information of the river basin from the point of view of potential floods. This task can be considered as one of the components of a more complex intelligent system for emergency management [7].


Last hour report

Significant discharge of the Casasola reservoir

The output flow of the Casasola reservoir is 650 m^3/seg which means an increase of 180 m^3/seg in the last hour. The normal output flow is 140 m3/seg. The current volume of the Casasola reservoir is 22 Hm^3 which means the 67% of its capacity. The maximum rainfall was located at Torcal with a value of 19 mm in the last 6 hours. Rainfall was also recorded at Colmenar (12 mm) and Casarabonela (9 mm) in the last 6 hours. According to the national weather forecast, significant rainfall is not expected for the next 24 hours in this area.

Figure 3: Example of summary in text mode.

In this context, information is received periodically at the control center about rainfall at certain locations, water levels and flow discharge in reservoirs and flows in certain river channels. A typical number of variables in a basin with a SAIH control center is about 500 measures every Δt (for example $\Delta t=30$ min). The analysis of a hydrologic situation requires usually data from the last 24 hours, so a typical amount of data about 24,000 quantitative values. As a result of the summarizing process, relevant states are reported such as important rainfall at certain location, a significant increase of flow at certain location or, during the evolution of a particular storm, significant decrease of rainfall or flow at certain locations.

Figure 4: Example of 3D animation on a virtual terrain presenting relevant information.

In order to present the summarized information, different modes have been considered such as text, 2D graphics and 3D animations on a virtual terrain. Figures 3 and 4 show examples of these types of presentations. To automatically construct the report, the computer system includes planner based on HTN (Hierarchical Task Networks) [8]. The planner follows a template-based strategy, with abstract presentation fragments corresponding to discourse patterns. According to the type of user the information is presented using different devices such as mobile phone (with *sms* messages), fax or a computer screen.

4 Summary and discussion

In summary, the paper describes our approach for a summarization problem. The problem is summarizing the behavior of a complex dynamic system, where partian and approximate knowledge about structure and behavior is available. The main contributions of our work for this problem are: (1) a notion of relevance based on the distance to management goals which provides a particular strategy for summarization, and (2) the identification and representation of different types of available knowledge together with an inference procedure. The approach presented in this paper has been initially validated in the domain of hydrology with successful preliminary results with partial models. Currently we are working in a more extensive and complete evaluation of the solution and its integration with presentation methods.

Our approach is related to several general AI fields such as event summarization, model-based problem-solving methods and relevance reasoning. Within the field of event summarization [1] [9], there are techniques that go from domain dependent approaches (taking into account saliency and abstraction techniques) to domain independent solutions based on statistic analysis. Our approach is a domain dependent approach that follows specific inference strategies derived from the context of surveillance of dynamic systems.

On the other hand, in the field of model-based solutions, our approach is related to modeling approaches for qualitative physics [10] [11] [12] such as CML [3] and DME [4]. These general approaches are theoretical solid approaches that in practice usually need to be formulated with additional control mechanisms to avoid computational problems. Our approach is not oriented for prediction nor for diagnosis so it follows a simpler and more efficient representation for the behavior that requires less knowledge acquisition effort. Compared to methods for diagnosis [13] our approach does not look for hidden causes starting from given symptoms. Instead, it selects and summarizes the relevant information in the measured data.

Relevance reasoning has been studied from different perspectives such as philosophical studies or logic-based formal systems [14]. In artificial intelligence it has been considered in different problems such as probabilistic reasoning [15] or knowledge base reformulation for efficient inference [16]. Closer to our approach, relevance reasoning has been used in the representation of dynamic systems. For example, relevance reasoning is applied in compositional modeling (dynamic selection of model fragments for simulation) [17] which is not the same task

performed by our method. Our approach is closer to the case of explanation generators of device systems such as the system of Gruber and Gautier [18]. As in our approach, this system defines relevance based on state transitions. However, our method includes additional domain dependent mechanisms for relevance based on the management strategy, a filtering procedure based on causal knowledge, and additional abstraction techniques based on hierarchies of components.

Our approach is also related to techniques for summarizing time series data. For example, our work presents certain commonalities with the SumTime project [19]. Compared to our work, this project pays more attention to the natural language generation from temporal series while our work is more centered on using a particular representation of the dynamic system that provides adequate solutions for data interpretation, aggregation and filtering.

Other similar systems but restricted to the field of meteorology have been developed for summarizing [20] [21] [22]. For example, the RAREAS system is a domain dependent application that generates text summaries of weather forecast from formatted data. In contrast, our method has been conceived in general to be used in different domains such as road traffic networks, water-supply distribution networks, etc.

Acknowledgements. The development of this research work was supported by the the Ministry of Education and Science of Spain within the E-VIRTUAL project (REN2003-09021-C03-02). In addition to that, the Ministry of Environment of Spain (*Dirección General del Agua*) provided information support about the domain in hydrology. The authors wish to thank Sandra Lima for her valuable comments and her work on the implementation of the method.

References

1. Maybury, M. T.: “Generating Summaries from Event Data”. Information Processing and Management: an International Journal. Volume 31. Issue 5. Special issue: Summarizing Text. Pages: 735 – 751. September 1995.
2. Schreiber G., Akkermans H., Anjewierden A., De Hoog R., Shadbolt N., Van de Velde W., Wielinga B.: “Knowledge engineering and management. The CommonKADS methodology” MIT Press, 2000.
3. Bobrow D., Falkenhainer B., Farquhar A., Fikes R., Forbus K.D., Gruber T.R., Iwasaki Y. , and Kuipers B.J.: “A compositional modeling language”. In Proceedings of the 10th International Workshop on Qualitative Reasoning about Physical Systems, pages 12-21, 1996.
4. Iwasaki Y. and Low C.: “Model Generation and Simulation of Device Behavior with Continuous and Discrete Changes”. Intelligent Systems Engineering, Vol. 1 No.2. 1993
5. Gruber T. R. and Olsen G. R.: “An Ontology for Engineering Mathematics”. In Jon Doyle, Piero Torasso, & Erik Sandewall, Eds., Fourth International Conference on Principles of Knowledge Representation and Reasoning, Gustav Stresemann Institut, Bonn, Germany, Morgan Kaufmann, 1994.
6. Borst P., Akkermans J. M., Pos A., Top J. L.: “The PhysSys ontology for physical systems”. In R. Bredeweg, editor, Working Papers Ninth International Workshop on Qualitative Reaso ning QR'95. Amsterdam, NL, May 16-19. 1995.

7. Molina M., Blasco G.: “A Multi-agent System for Emergency Decision Support”. Proc. Fourth International Conference on Intelligent Data Engineering and Automated Learning, IDEAL 03. Lecture Notes in Computer Science. Springer. Hong Kong, 2003.
8. Ghallab M., Nau D., Traverso P.: “Automated Planning: Theory and Practice”. Morgan Kaufmann, 2004.
9. Maybury, M. T.: “Automated Event Summarization Techniques”. In B. Endres-Niggemeyer, J. Hobbs, and K. Sparck Jones editions, Workshop on Summarising Text for Intelligent Communication. Dagstuhl Seminar Report (9350). Dagstuhl, Germany. 1993.
10. Forbus K. D.: “Qualitative Process Theory”. Artificial Intelligence, 24: 85-168. 1984.
11. de Kleer, J., Brown, J.: “A Qualitative Physics Based on Confluences”. Artificial Intelligence. 24:7-83. 1984.
12. Kuipers B.: **“**Qualitative simulation**”.** Robert A. Meyers, Editor-in-Chief, Encyclopedia of Physical Science and Technology, Third Edition, NY: Academic Press, pages 287-300. 2001.
13. Benjamins R.: “Problem-solving methods for diagnosis”. PhD thesis, University of Amsterdam, Amsterdam, The Netherlands. 1993.
14. Avron A.: “Whither relevance logic?”. Journal of Philosophical Logic, 21:243-281. 1992.
15. Darwiche, A.: “A logical notion of conditional independence”. Proceedings of the AAAI Fall Symposium on Relevance, pp. 36-40, 1994.
16. Levy A., Fikes R., Sagiv, Y.: “Speeding up inferences using relevance reasoning: a formalism and algorithms”. Artificial Intelligence, v.97 n.1-2, p.83-136, Dec. 1997
17. Levy A.. Iwasaki Y., and Fikes R.: “Automated Model Selection Based on Relevance Reasoning”, Technical Report, KSL-95-76, Knowledge Systems Laboratory, Stanford University. 1995.
18. Gruber, T. R., Gautier, P. O.: “Machine-generated Explanations of Engineering Models: A Compositional Modeling Approach”. Proceedings of the 13th. International Joint Conference on Artificial Intelligence. 1993.
19. Sripada, S. G., Reiter, E., Hunter, J., and Yu, J., "Generating English Summaries of Time Series Data Using the Gricean Maxims", Proceedings of the 9th International Conference on Knowledge Discovery and Data Mining SIGKDD, Washington, D.C. USA, 2003.
20. Kittredge R., Polguere A., Goldberg E.: “Synthesizing weather forecasts from formatted data”. In Proc. International Conference COLING-86, Bonn, August 1986.
21. Bourbeau L., Carcagno D., Goldberg E., Kittredge R., Polguere A.: “Synthesizing Weather Forecasts in an Operational Environment”. In Proc. International Conference COLING-90, vol. 3, 318-320, Helsinki, August 1990.
22. Goldberg, E.; Driedger, N.; Kittredge, R.I.: Using natural language processing to produce weather forecast”. IEEE Intelligent Systems and Their Applications. Volume 9, Issue 2, April 1994.