COLING 82, J. Horecký (ed.)
North-Holland Publishing Company

RECOGNITION OF ABSTRACT OBJECTS - A DECISION THEORY APPROACH WITHIN NATURAL LANGUAGE PROCESSING

Gerhard Knorz

Fachbereich Informatik, FG Datenverwaltungssysteme II
Technische Hochschule Darmstadt
Karolinenplatz 5
D-6100 Darmstadt
W-Germany

The DAISY/ALIBABA-system developed within the WAI-project represents both a specific solution to the automatic indexing problem and a general framework for problems in the field of natural language processing, characterized by fuzziness and uncertainty. The WAI approach to the indexing problem has already been published [3], [5]. This paper however presents the underlying paradigm of recognizing abstract objects. The basic concepts are described, including the decision theory approach used for recognition.

1 THE "WAI" AND THE "AIR" PROJECT[1]

The DAISY/ALIBABA system [1], [2], [3], as developed at the Technical University Darmstadt analyses abstracts and describes them according to the coordinate indexing philosophy using a prescribed set of descriptors. To perform this task, a domain dependent dictionary is needed. Estimating the non-existence of suitably sized dictionaries to be one of the main problems for research and development of automatic indexing [4], in 1978 the WAI project started with dictionary construction. The two completed dictionaries are

- FST, covering the scope of food science and technology[3] and
- PHYS, covering the scope of Physics, a part of INIS (International Nuclear Information System)[4].

Different procedures for generating dictionary data were developed and applied. To classify them and to unify the created data is one of the main tasks of dictionary construction (described in detail in [3], [4]). This cannot be done without examination of their influence on the quality of the resulting indexing. To perform indexing tests, the development of DAISY and ALIBABA was another important objective of WAI.

Indexing results are reported in [4], [5], [6] which are based on consistency tests only, using the manual indexing as a standard. To confirm or to modify these results, the AIR project is now preparing a retrieval test on the physics data base INKA-PHYS of the Fachinformationszentrum FIZ 4 (Energie, Physik, Mathematik; Karlsruhe) (order of magnitude: 10.000 documents, 200 search requests). The indexing will be based upon the new dictionary PHYS-2 which is to be constructed using about 80.000 documents of the INKA-PHYS data base.

2 THE BASIC PRINCIPLES UNDERLYING THE "WAI"/"AIR" APPROACH

The WAI/AIR approach represents both a specific solution of the indexing problem

and a general framework for a wide class of problems within natural language processing and other fields.

This paper will only give reference to details of the particular solution published elsewhere. The objective of this work is to present the general framework derivable from the basic principles underlying the WAI and the AIR project:

(1) Knowledge bases are very important for problem solving. But to presuppose knowledge for an automatic system must not question its applicability, caused by non-existent procedures for construction of knowledge bases of an indispensable size. The realistic appropriate solution is the main aim rather than a perfect one.

(2) Controlling the quality and expenditure of effort of a system must not wait until it is put into practice. System development has to be guided by a control derivable from the task to be performed.

(3) The algorithms that make the bases of the procedure should not be assumed to be perfect. Applied to complex tasks, it is a fundamental fact that they are based on simplified models.

The principles can be considered to be a guideline for designing application oriented systems. With good reason it is claimed that the quality of such a system can be determined by evaluation in application environments only (see for example [7], [8]). This cannot be done without empirical studies of the user-system interaction.

The paradigm of recognizing abstract objects presented here is an approach to integrate the evaluation aspect into system development. It is also an approach to problems, for which no perfect solutions exist or seem to be applicable.

3 RECOGNITION OF ABSTRACT OBJECTS

3.1 THE DEFINITION OF THE RECOGNITION TASK

The basic idea is to use the application environment itself to get an implicit description of the problem. Whenever talking about a particular application environment there is no other way then to take a conceptual model M_E as a basis which determines the adequate concepts (see [9], or see also [10][5]).

Here, a conceptual model has to be formulated in this way, that it defines (abstract) objects $(\tilde{x},k)$, $\tilde{x}\epsilon X$, $k\epsilon K$. $\tilde{x}$ denotes those aspects of an object which can be observed directly with regard to the problem, K denotes a set of object classes. A model m_E of the application environment gives an implicit definition of the (recognition) problem, by forming a continuous stream of abstracts objects.

To develop a recognition system (RS) is nothing more than the finding of a suitable mapping $e: \tilde{x} \rightarrow e(\tilde{x})$ that recognizes an actual $\tilde{x}$ to be $(\tilde{x}, e(\tilde{x}))$.

If the RS-m_E interface is identical to the system-user interface, then m_E may refer to the user's judgement directly, to define the co-occurrence of $\tilde{x}$ and k.

This is also adequate, whenever human cognitive capabilities are to be simulated. We give some examples:

- Information retrieval can be based upon recognition of document-query relationships (described in [6]). $\tilde{x}$ can be represented by (d,f) where d denotes the document, f denotes the query. k may be in the most simple case a member of the set {is relevant, is not relevant}, refering to the user's judgement.
- Expressions, possibly within the scope of a quantifier as well as hypotheses for inferences, can both be regarded as abstract objects. Determining the scope of a quantifier or drawing inferences can be based on the recognition of those objects by simulating human decisions.

Two other examples are given - avoiding the simulation approach:

- Complex tasks often require the testing of many hypotheses, which can be regarded as abstract objects. m_E may refer to the final results of the processing.
- In [6] a decision theory approach to optimal retrieval forms a basis for m_E, defining the task of indexing as recognition of document-descriptor relationships.

3.2 STRUCTURE OF THE RECOGNITION SYSTEM

The structure of the recognition system as presented here makes evident that the recognition problem arises essentially at the interface of two models:

- The (external) conceptual model M_E defining the recognition problem.
- The (internal) conceptual model M_I used to describe the object with respect to the recognition task.

M_I is part of the recognition system (Figure 1). It structures the object using the knowledge base, so that all available aspects that may influence the decision of the RS are included. In many cases it also initiates the recognition process, i.e. it constructs the hypothesis, represented by the object.

According to M_I a formal description x of $\tilde{x}$ is produced. We do not consider here the nature of M_I, that can be a sophisticated one with a strong theoretical foundation as well as a rather simple and heuristic one. Different models M_I might cause quite different recognition systems for the same task. The main point is, that M_I leads to an object description instead of a decision. Another point is, that both models M_E and M_I are essentially independent. This fact causes every

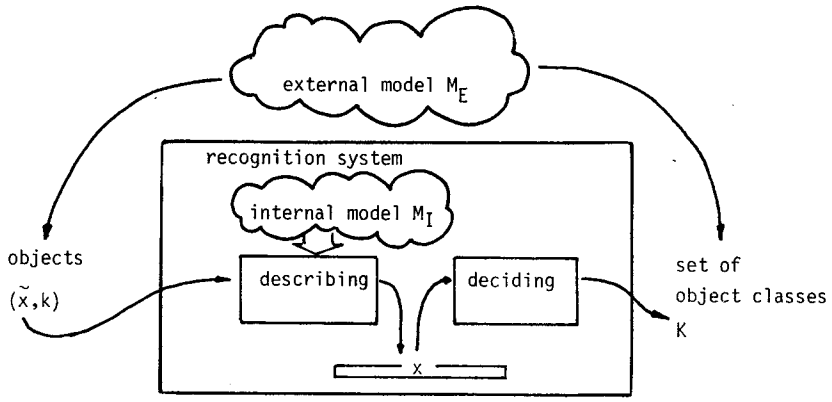

Figure 1 The recognition system and its environment

system RS^{M_I} - provided it is a deterministic one - to make incorrect decisions in some cases. That means, an 'optimal recognition systems' cannot be defined without taking the number of cases causing faults into consideration or - more precisely - the statistical properties of the application environment represented by m_E. The desision theory approach appropriate to the given situation is described in [5] and [6] with respect to the indexing problem. The approach requires that every single decision of RS is classified. This task is for the most part anticipated by M_E, which defines the set of object classes K. K determines the scope of possible faults. Those can be weighted independently by a loss function c: $(e(\tilde{x}),k) \rightarrow w$. With the model m_E given, a particular recognition system will cause an expected value E(w). The optimal system $RS^{M_I}_{opt}$ is the result of searching for this RS^{M_I} which minimizes E(w). It can be shown that the optimal decision $RS^{M_I}_{opt}(\tilde{x})$ can be based on the restricted probabilities $p(k|x)$. The mappings $e_k(x) = p(k|x)$ can be approximated by polynomial functions to be constructed automatically using a sample of objects $(\tilde{x},k)$. This way has been choosen by the ALIBABA system, that uses polynomial classifiers, adapted in the mean square sense [11] . The indexing results in [5] and [6] demonstrate that - applied to the indexing problem - the recognition approach and in particular the method of approximation is adequate for the problem.

4 DISCUSSION

The approach of recognizing abstract objects is evaluated using the paradig of automatic indexing. The model M_E refers - for practical reasons - not to the

retrieval process but to the decisions of human indexers. If a consistency factor (comparing manual and automatic indexing) measures the quality of automatic indexing, the set K requires two elements only. If a more sophisticated evaluation is intended, the set K can be increased, according to the kind of faults that should be considered. The classification of faults can for example depend on the descriptor under consideration.

For the model M_I used see for example [5] and [12].

We summarize the essentials of the suggested approach (the first point refers in particular to the indexing paradigm).

- The recognition problem causes one to regard two independent models: one with respect to retrieval and one with respect to analysis of abstracts. This point of view is important for an approach to optimal indexing [6], but it is not self-obvious. In [14] the retrieval oriented approach of Robertson and the indexing oriented approach of Harter [13] are brought together. The result is a one model approach like also other approaches in this field (for example [15]).
- The internal model M_I is restricted to the base of the decision to be made. This fact makes it very easy to additionally include a lot of knowledge and heuristic procedures, that might play a role only for decision making. There is no risc of causing faults by determining how to compute the decision, using this knowledge. Artificial intelligence approaches use a correspondant model M_I to determine the decision [16].
- The need for a model m_E implies an educational aspect with respect to evaluation. m_E ensures, that the gap between the optimal system $RS^{M_I}_{opt}$ and the ideal system (equivalent to m_E) is under control.

FOOTNOTES

1 WAI means Wörterbuchentwicklung für automatisches Indexing (dictionary construction for automatic indexing), [3]. The research was supported by the BMFT contract PT 131.05 to Technische Hochschule Darmstadt (march 1, 1978 - december 12, 1981).

AIR means Weiterentwicklung der automatischen Indexierung und des Information Retrieval (further development of automatic indexing and information retrieval). Supported by the BMFT contract PT 131.10 to Technische Hochschule Darmstadt (march 1, 1981 - december 31, 1983).

2 The order of magnitude of the two dictionaries may be characterized as follows: about 13.000 single words, 20.000 phrases and 100.000 term-descriptor relations each.

[3] The two volumes 3 and 4 of the abstract journal Food Science and Technology Abstracts (FSTA 71/72) containing about 33.000 documents were used as a basis for dictionary construction.

[4] The scope of Physics (INIS) is represented by about 40.000 documents.

[5] In [10] the term paradigm is used instead of 'conceptual model' that is taken here from [9].

REFERENCES

[1] Putze-Meier, G., DAISY - Darmstädter Indexierungssystem, to appear as a report, Technische Hochschule Darmstadt, Fachbereich Informatik, DVS II (1982).

[2] Knorz, G., Softwaresystem ALIBABA, Adaptives lernstichprobenorientiertes Indexierungssystem, basierend auf Beschreibungen abstrakter Objekte, Bericht DV II 82-1, Technische Hochschule Darmstadt, FB Informatik, FG DVS II, (1982).

[3] Lustig, G., Das Projekt WAI: Wörterbuchentwicklung für automatisches Indexing, to appear in the proceedings of the Deutscher Dokumentartag 1981 (Saur KG, München, 1982).

[4] Lustig, G., Über die Entwicklung eines automatischen Indexierungssystems, in: Krallmann, D. (ed.), Dialogsysteme und Textverarbeitung (LDV-Fittings, Essen, 1980).

[5] Knorz, G., Automatic Indexing as an Application of Pattern Recognition Methods to Document-Descriptor Relationship, applied informatics 1 (1982) 1-10.

[6] Knorz, G., A Decision Theory Approach to Optimal Automatic Indexing, to appear in the proceedings of the GI/ACM/BCS Conference (Berlin, May 1982).

[7] Krause, J., Lehmann, H., User Speciality Languages. A natural language based information system and its evaluation, in: Krallmann, D. (ed.), Dialogsysteme und Textverarbeitung (LDV-Fittings, Essen, 1980).

[8] Ackermann, Ammon, Ebert, Krause, Krug, Marschke, Sauerer, Zimmermann (ed.), Cobis. Computergestütztes Büro-Informationssystem als Pilotanwendung von CONDOR, BMFT-report (Karlsruhe, 1982).

[9] Schmitt, B., Computer Science and the General Theory of Models - An Introduction, applied informatics 1 (1982), 35-42.

[10] Kuhn, T.S., The structure of Scientific Revolutions. (Chicago, 1970).

[11] Schürmann, J., Polynomklassifikatoren für die Zeichenerkennung - Ansatz, Adaption, Anwendung -, (Oldenbourg Verlag, München, 1977).

[12] Knorz, G., Mustererkennung im Bereich der inhaltlichen Erschließung von Texten, in: Radig, B. (ed.), Modelle und Strukturen (Springer Verlag, Berlin Heidelberg New York, 1981).

[13] Harter, S.P., A probabilistic approach to automatic kexword indexing. Part I: On the distribution of speciality words in a technical literature, Journal of the ASIS, 26 (1975), 197-206, Part II: An alogorithm for probabilistic indexing, Journal of the ASIS,26 (1975) 280-289.

[14] Robertson, S.E., van Rijsbergen, C.J., Porter, M.F., Probabilistic models of indexing and searching, in Oddy, R.N., Robertson, S.E., van Rijsbergen, G.J., Williams,P.W. (ed.), Information Retrieval Research,(Butterworth,London,1981).

[15] Cooper, W.S., Maron, M.E., Foundation of Probabilistic and Utility-Theoretic Indexing, IACM, 1/25 (1978) 67-80.

[16] Wahlster,W., Implementing Fuzziness in Dialogue Systems, in Rieger, B. (ed.) Empirical Semantics, (Brockmeyer, Bochum, 1981).