Technical Report

UCAM-CL-TR-211
ISSN 1476-2986

Number 211

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

The role of artificial intelligence in information retrieval

Karen Spärck Jones

November 1990

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500
https://www.cl.cam.ac.uk/

Technical reports published by the University of Cambridge
Computer Laboratory are freely available via the Internet:

https://www.cl.cam.ac.uk/techreports/

ISSN 1476-2986

THE ROLE OF ARTIFICIAL INTELLIGENCE

IN

INFORMATION RETRIEVAL

Karen Sparck Jones

Computer Laboratory, University of Cambridge
New Museums Site, Pembroke Street, Cambridge CB2 3QG, UK

November 1990

Abstract

This paper reviews four potential roles for artificial intelligence in information retrieval, evaluating AI from a realistic point of view and within a wide information management context. The conclusion is that AI has limited potential, not just because AI is itself insufficiently developed, but because many information management tasks are properly shallow information processing ones. There is nevertheless an important place for specific applications of AI or AI-derived technology when particular constraints can be placed on the information management tasks involved.

Note

This paper was originally presented as the lecture given for the ASIS New Jersey Chapter Distinguished Lectureship Award, 1990.

THE ROLE OF ARTIFICIAL INTELLIGENCE IN INFORMATION RETRIEVAL

Karen Sparck Jones

Computer Laboratory, University of Cambridge
New Musuems Site, Pembroke Street, Cambridge CB2 3QG, UK

INTRODUCTION

Artificial intelligence (AI) has been established as a research area for some thirty five years, at least since the Darmouth Conference of 1956; and fully operational, commercially valuable systems exploiting AI have existed for a decade, notably in the area of expert systems, for example DEC's system for choosing and configuring computers. It has been suggested, with increasing frequency, that we can have intelligent information systems, meaning, broadly, systems able to provide effective access to information depending on texts or documents through the autonomous, adaptive manipulation of concepts. It is important that those engaged in IR should not be abused by the improper use of the word "intelligent" and more importantly, be misled by ignorance of the real state of AI or by misconceived goals based on a failure to recognise the essential properties of IR, interpreting this as document indexing and searching in the widest sense. My intention is to look analytically at the potential roles for AI in IR, and hence realistically at what we can expect, operationally, in the next decade. Thus I think it is important not to adopt uncritically the notion that as information management and, crucially, the management of linguistically-conveyed information, clearly requires human intelligence, we can just aim, admittedly as very long term goal, to replace this by machine intelligence.

To understand the true problematic here, we need to appreciate the properties of documentary information systems and, in the context of realistic operational aims, the assumptions of AI, especially as currently viewed.

Artificial intelligence

Thus I shall start by noting what, reflecting the present general view, is regarded as defining AI. This is the use of computers to carry out tasks requiring reasoning on world knowledge, as exemplified by giving cooperative responses to questions in a situation where one is dealing with only partial knowledge and with indirect connectivity. For example at a station we might inquire "What platform for the London train?" and get the response "The 11.30 has been cancelled. There has been a derailment at Hitchin. The 12.30 will probably be cancelled too." This response depends on all kinds of knowledge about trains and follows extensive inference, for instance if someone asks about the London train it means they want to go to London as soon as possible, and the next train for London is the 11.30. But

this is not certain knowledge (they could be wanting the next train from London), and there is equally a quite long and elaborate chain of reasoning between the question and the remark about the 12.30 probably being cancelled too, because derailments normally block lines for a long time. All this illustrates the kind of complex reasoning on necessarily imperfect knowledge of the sort AI seeks to provide.

There are two particular points about the example worth noting. The first is that much of the knowledge in the base is generic, e.g. that derailments cause extensive disruption, so what the system has to do is deploy this generic knowledge in relation to specific instance data in order to make relevant specific inferences, in this case about the 12.30. The second is that it is important to recognise goals and plans, and the beliefs on which these are based, e.g. the inquirer's goal is to get a London train he believes exists. All of these things seem obvious to us, but they have to be made explicit, in fine detail, for a system to reason effectively.

With more exigent applications than train inquiries, one can have a very extensive knowledge base and reasoning, and the real problem is that of capturing and representing the necessary knowledge, given that so much of it is approximate: for example, if the user asks about the next fast train, what is a fast train? These problems are the same, moreover, whether one is dealing with naturally interactive applications, or with ones where natural language is a means of communication but the system's task is not a naturally interactive one, or with applications where language is a prime source of information but is not necessarily exploited interactively.

What, then, are the apparent roles for AI in IR?

POTENTIAL ROLES FOR ARTIFICIAL INTELLIGENCE

It is possible to distinguish four generic roles. First, in the central documentation activity, information characterisation, i.e. in subject indexing and matching, where the claim is that the ability to automatically identify document or text concepts and relations requires AI depending on natural language undertstanding. The second role is in the complementary search formulation of information seeking, i.e. in identifying the user's need and in expressing this as a document request and eventually as a search statement. It is proper to separate these two roles because though they have a common characterisation step, indexing, for both documents and requests, there is an important difference, from the AI point of view, between passive documents and active searches, and between independence of, and dependence on, the individual user. There is a third possible role for AI in other, supporting, information management operations, like cataloguing, or notification. Finally, while the tacit assumption so far has been that we are dealing with a single document collection, it has been suggested we should look to AI in a fourth role not just even for multifunctional use of a single collection, but to support the use in one or more ways of different, perhaps of very different types of, collection, in systems integration.

I want, now, to tool more carefully at what is involved in claiming these roles for AI.

Information characterisation

In its strongest form, the claim for AI here is that given a set of documents, what we should do is build a knowledge base from them; that is, we should extract their information content and integrate this into a single whole. The reason for doing this is that we remove the present indirection in retrieval: the user wants to find something out, and we allow him to do this directly, by addressing the knowledge base; he does not have to work indirectly, reaching documents he has to read for their content, via their index descriptions.

So, for example, if we are in the area of medical research and have a question, 'Does tungsten deficiency cause Smith's Disease?',[1] we would have a knowledge base built from medical papers with information couched in, say, some logical representation language about diet, blood, metabolism, etc, and about the properties of Smith's Disease, along with an inference apparatus for reasoning over both specific and general relationships between e.g. diet and blood. We could then infer, if tungsten is a mineral and metabolic processes transfer diet components to blood constituents, and Smith's Disease affects liver metabolism for minerals, that tungsten deficiency in the blood is a symptom and not a cause of Smith's Disease.

Thus the model being proposed for information searching is of question answering, and information characterisation is designed to support this. The user has a question to which he wants an answer, which reasoning on a knowledge base would supply. Document retrieval is thus simply a regrettable substitute for this proper process.

I believe this is a fundamentally mistaken view (quite apart from any questions of its current feasibility). It assumes, in replacing document texts in a natural language by a knowledge base in an artificial language, and specifically in a necessarily artificial language, that there is no loss of meaning. But the way things are expressed in natural language is informative in itself. The individual words in a text, and their ordering, are part of the content of that text. The model just outlined also assumes that the user wants to find out something, when he rather wants to find out about something. Finally, it is not clear how the information reduction needed to allow access to large quantities of information, currently supplied for text collections by index descriptions, is to be achieved.

This model embodies a very strong version of the AI claim. There is a weaker version which accepts that the document texts themselves are ultimately needed, but claims that we will still get better access, i.e. characterisation and matching, via an integrated knowledge base which embeds some text information and is exploited inferentially to reach the right points linking to the documents themselves. Thus the

[1] This example in all its ramifications is imaginary, and is deliberately designed to avoid issues of specific scientific plausibility.

base could contain some facts about deficiency and illness, for example, and about Smith's Disease, so reasoning would indicate that the relationship between mineral deficiency and Smith's Disease could be either cause or effect and would point to a document discussing which in detail. A knowledge base of this sort, which would be a single whole, would be a structurally more complex and formally more powerful system than e.g. a conventional classification: it could, for instance, allow role relationships and inheritance on these as well as by straight set membership.

This approach would avoid some of the previous problems, notably those of expression and of reduction. But quite apart, again, from questions of feasibility in constructing the base, and especially of deriving the base in large part from the documents, there is a real problem in the linkage of base and documents, with their differences of language, style and grain. There is a particular problem of providing discriminating links to distinct documents. Linking is difficult enough with conventional classifications, but succeeds because the links are recognised to be coarse. Here something better is needed, or the supposed gains from reasoning on the base will be lost.

But the main problem with this model is that the mix of question answering and non question answering now achieved seems very strange and unmotivated: how do we decide which is to be which?

The third, even weaker, version of the AI claim therefore abandons the unified knowledge base and inference on it, but maintains that AI is still needed to derive proper individual document characterisations, i.e. to capture and represent their key content, and to allow proper matching. This is using natural language understanding to obtain descriptions in some unambiguous and normalising meaning representation language, say with predicate-argument logical forms. So, for example, with a description of Document 100 in the form of two propositions about Smith's disease and metabolism and about liver and tungsten, we could do inferential matching on these for a query proposition about tungsten and Smith's Disease, i.e. we could relate a single query proposition in a well-organised way to the conjunction of two document propositions.

But unfortunately, while these AI proposals may look exciting compared with present automatic indexing and matching using text keywords, they are just proposals for obtaining older-fashioned types of complex indexing description involving semantic and/or syntactic control hitherto provided manually, and referred to as subject headings and so forth. They are certainly generically the same, and may even be so in detail. And setting aside the challenge of deriving these descriptions automatically, there is not enough evidence from such experiments as have been done with sophisticated manual descriptions of this sort, to show that they give a better retrieval performance than much cruder ones using natural language and coordination. The experiments that have been done are not ideal, but there is no evidence that sophisticated descriptions of this sort pay off better than the best natural language strategies based on coordination, mixes of substitute and additional terms, with terms embodying but not indicating different relations, term weighting and so forth.

The final, weakest, position is that we may still want AI to select good natural language index terms, and particularly compound terms encapsulating significant meaning relationships not necessarily figuring in a consistent or transparent way in texts. So, for example, we could analyse a document or request text containing the phrase "metabolism within the liver" to recognise the compound concept 'liver metabolism'. The assumption is that simple coordination on words is not a good enough way of signalling conceptual units, but that coordination on compound units gives the right level of descriptive constraint, more than single words, but less than subject headings.

In principle doing this properly, and even more identifying key ones, requires natural language understanding based on world knowledge. But in practice it may be possible to do it sufficiently well on a more narrowly linguistic basis. This is something which has not necessarily been done before manually in the specific forms suggested, or been thoroughly evaluated, so experiment is needed. But as it may be not only necessary but sufficient to do it with restricted language processing rather than full language understanding, this means that it can hardly be labelled a real use of AI.

My analysis of information characterisation thus leads to an essentially negative conclusion about AI, namely that it has no potential role in indexing and matching because these do not extract or use world knowledge directly. But it is most important to emphasise that this is only for the general case where we have large scale and variety of subject on the document side, and variety of need on the user side. We may have a quite different situation in specialised contexts where an AI approach, in any of the first three guises, may be both justifiable and feasible. I shall return to this possibility later.

Information seeking

The second major area where it has been suggested AI has a role is in search formulation, i.e. in the development and characterisation of the user's request, not in the actual matching of the request against the collection.

The notion here is of automating the intermediary, to give the user accessing information a helpful front end in the form of an expert system exploiting knowledge about users, subjects, literatures, document descriptions and search techniques in order to determine the user's real need and how one may seek to satisfy it, of course via dialogue and also iterative searching.

So for instance in the medical case we might have a front end with knowledge about doctors versus medical students, about illness and health in general, even considerable though non-specialist medical knowledge (more than a human intermediary), about medical literature, services like Medline and indexing languages like MeSH, and about search strategies like Boolean disjunction. We would then have rules so, for example, given a student user and a general topic request, we would look for a survey document.

There is no doubt that human expert help is valuable, or that embodying it automatically requires AI: the front end system would certainly need to reason on knowledge in the manner defined earlier.

But we have only to think about the full implications of attempting to build such an expert intermediary to see how far we are from being able to do this. As in a serious attempt to derive a knowledge base from documents, the automated intermediary depends on language understanding and has the fundamentally hard task of determining what linguistically-conveyed information adds up to. But we also have the additional factor of dynamic interaction with the user: the user is not passive like a document being indexed. Although language is the essential means of communication, so we need language understanding for this, the fact that we have a more complex, multifacetted task involving finding out about the user, about all aspects of his need including the background context, encompassing purpose, and so forth, as well as about the topic itself, and also e.g. the form of material to be output, means we will need a much more extensive knowledge base and reasoning apparatus than in the previous case. It is clear that where the AI approach to information characterisation is misconceived for the general case, the problem here is that attempting to replace the human intermediary for the general need is quite unrealistic. It may however, as for the previous role, be quite reasonable to seek intelligent interfaces for more specialised or limited contexts where natural restrictions on the task make it feasible to apply current or foreseeable AI, and some expert system front ends, as illustrated later, have already been built.

Support functions

Outside the central area of indexing and retrieval concerned with topic and content there are other document and information management functions in which AI may also have a role. Some of these tasks, like abstracting, if done properly and we also knew how to do them automatically, would certainly require natural language understanding and thus AI. With other content-based tasks like alerting this is less clear.

But there are also functions not involving or confined to text content which are very important, and require considerable experience and skill, like cataloguing. It has been suggested that we may be able to build expert systems applying knowledge-based rules for this, if not to replace the librarian altogether at least to support him, say by prompting and suggestion, so for instance we might have a rule suggesting language of publication from place of publication. Again, it is certainly reasonable to attempt this as a potentially sufficiently restricted, if not trivial, task. Another possibility would be choosing services or databases in online searching.

Systems integration

I have been considering the potential for AI so far as if we had to deal only with a single collection or single task independent of others, and with real or logical objects like document texts or catalogue entries in quite distinct use and time contexts. But the same objects can of course be of interest from many different

points of view at the same time, and equally the same person can be interested in many different objects in different ways for different purposes at the same time: for example, a document for reading or quoting, a catalogue entry for subject search or for locating an item; or, for instance, rereading a text and checking a bibliographic item at the same time, during document preparation.

With automation we have an increasing provision of different kinds of object and function through machines, and an increasing interest in integrating them for the user within a single computational system or environment. For example, in the wider medical context we might have collections of scientific articles, standard data tables, clinical and organisational material including both patient records and administrative items of all kinds, and laboratory test files; and insofar as all these are associated with individual institutions, we may also have national or international resources on the one hand and the individual's personal files on the other containing, for example, notes and email messages. We would also have second order objects like directories as well as first order ones. (This variety refers just to language based items: there might be graphical ones as well.)

Any individual might engage in different information management activities aimed not only at picking up existing information by, for instance, looking for articles, checking references, inspecting records, etc, but also at putting down new information through writing, glossing, assembling or extracting, and so forth. We may further have quite different classes of users, e.g. administrators as well as researchers, using some or all of the material in the same or different ways, for example the same patient record may be used on the one hand for investigating a disease and on the other for scheduling hospitalisation.

The heady integration scenario is that given, say, a researcher writing a paper comparing tungsten and bismuth deficiency in relation to liver disease and asking for a literature search on this topic, the system would also automatically, i.e. autonomously, check the data tables for tungsten and bismuth, note the chemical group relationship of bismuth with arsenic, pick up the fact that arsenic is poisonous, and stimulate a search of patient records for arsenic poisoning and liver disease, sending a message at the same time to a specialist for any experience of bismuth poisoning. In parallel, in this multi-user system, we might have a clinical practitioner with a liver patient being led via liver disease and tungsten deficiency to recorded tests and a specialist clinic appointment for the patient, while an epidemiologist noting an increased incidence of liver disease might be led through blood to diet to minerals in vegetables, and from there to scanning patient records for vegetarianism, and to a library search for tungsten in water supplies.

The assumption is that we can only achieve this kind of multi-resource, multi-use integration effectively through AI, since a substantive knowledge base and reasoning infrastructure is required to connect different points in the whole and to lead each user to the objects and activities he needs, i.e. to give the individual a virtual personal information system as a dynamic construct within the larger impersonal one. And indeed we can see from the example that it would be hard to get a system of this kind without AI.

There is no doubt that an integrated system adapting itself to each and every user is a very attractive idea, but quite apart from any questions about current feasibility, there is a more serious question about the sense of the whole idea.

There would in fact be so much heterogeneity in the system, heterogeneity of objects, functions, and satisfaction relations, that the suggestion that integration can be achieved via a single common propositional knowledge base and reasoning engine seems fundamentally misconceived. The knowledge base would have to be a sufficiently comprehensive and discriminating infrastructure for things as different as papers and database lines, as books and invoices, as first order messages and second order indexes, and for objects or object sets as different in grain size as the individual's electronic mail for the day or personal current files and the whole of Medline. How can one connectibly describe such varied things so as to be able move appropriately from one to the other?

Functions like writing a paper or citing it are very different, and the scenario for writing a paper and the description for retrieving it when written may also be quite different. Relevance relations, in the necessarily broadest sense, may equally be quite different, as in extracting a known catalogue entry compared with assessing topic papers. How could one bring the necessary knowledge about all these various kinds of object, and the variegated functions on, and satisfaction relations between, them together in a single common knowledge base in some unambiguous meaning representation language independent of ordinary language, and carry out all the functions and satisfy all the relations effectively through a common inference engine working on the base? How in particular could one do this given the fact that the system by definition is designed to link the public with the private, and the more wide ranging the public and hence miscellaneous its members, the less likely it is one could build a common base which was concrete and finely-structured enough to meet each of their separate individual requirements? There are well known problems about comprehensive thesauri and library classifications as effective common resources, and they aim at only a tiny fraction of what is required here.

This heterogeneity of object, function and relation is compounded by the fundamental imprecision of many information management activities, that is by imprecision due to the necessary variability of natural language, to indirectness of access, and to indefiniteness of need. This suggests that we can only achieve the required content connectivity with the much looser, indicative linkage given by a word network, in the broadest sense of the term. We can use natural language words (or stems or components or whatever) as connective keys because there is some community of meaning in the different uses of a word, however they are particularly intended, especially if we also offset ambiguity by redundancy, i.e. by working with extensive sets of words. The connectivity that can be achieved through a lexical network will not be very strong: but it is all that is available.

This connectivity is also stronger than it may appear. Thus it is important to recognise that the network, though fundamentally simple, can cover not only different kinds of primitive or basic units, but can capture all kinds of relations, syntagmatic as well as paradigmatic, and can therefore generate larger-scale units

than the initial minimal lexical ones. Thus we can expect to capture syntagmatic, compound concept, units in this way, though we may if we wish have multi-word primitive units where these are appropriate. The important point is that the relations in the network, as semantic relations, are implicit and informal rather than explicit and formal, though we can invoke the latter as expressed in an independent terminological knowledge base like a thesaurus if we want to, as a supplementary aid to our network manipulation.

We have still to rely heavily on the user's own understanding of his natural language. But we should help him by exploiting all the straightforward computational technology we can, especially at the interface, through the use of windows and so forth. Hypertext may similarly be a valuable support. It is a mistake to regard hypertext as embodying autonomous intelligence and therefore as being a form of AI, but this is not to deny its ancillary utility.

Within this framework it is quite proper to think of having AI-dependent embedded component systems for particular purposes, as illustrated later, though we cannot expect to be able to extend these indefinitely. We may also access these, when we do not know how to enter them directly, through the simple natural language associative network, since their terminology will, where it is to be comprehensible to the user, tie in with the rest of the system's collective language apparatus.

CONCLUSION ON ARTIFICIAL INTELLIGENCE

The overall conclusion about AI in IR, and more broadly in text-related information management, is thus that it is essential not to overestimate the potential for AI, for two reasons. The first is that where really hard tasks, depending on extensive world knowledge and on deploying this in a flexible and constructive way, like providing advice including explanations, are concerned, we will not be able to replace humans by mcahines in the foreseeable future. The second, very different reason is that there are many information operations which, even when done by people, are properly of a rather shallow, linguistically-based kind, which do not involve elaborate reasoning or complex knowledge. So here AI, whether feasible or not, is not only not needed in any strong form: it is fundamentally inappropriate.

But while I do not believe we should look to a central position for AI, I think it has a very valuable contribution to make. It can contribute, first, with specialised systems or subsystems, including those for sub-functions, where the knowledge base or inference can be limited, either intrinsically or because the user can complement the limitations. The second potential contribution is via techniques derived from AI, though these are not applied within a full AI context, for instance techniques found in rule based systems or in natural language processing. AI in either of these forms can be set within the framework of more or less any mix, as suited to circumstances, of conventional or unconventional information management, of unified or separate systems, or of low or high user participation. So there may be a useful place for some modest AI to support the overworked professional or lonely user in less rich

environments than those for which AI is being developed.

I shall illustrate these more modest, potentially valuable uses for AI or its descendants with three examples.

Example 1

The first example is information characterisation using natural language processing, as investigated by Sparck Jones and Tait, i.e. a use of AI-derived technology. The approach here is designed to analyse sentences syntactically and semantically to build a meaning representation with propositional components, from which more or less extensive constituents (predicates and case-related arguments) can be extracted. We then take these conceptual index term sources and generate syntactic and semantic variants according to allowed rules and classes, eventually providing alternative surface forms for the underlying concepts. If we process requests in this way we can create sets of terms (to be used within an overall Boolean coordination structure) for searching document (or abstract) texts.

This strategy requires a non-trivial lexicon and syntactic and semantic rules, including ones supplying prepositional phrases as alternatives to compound nouns; but it does not require a base of world knowledge. The work done on it so far, however, has only been in the form of laboratory experiments testing feasibility, and it needs much more investigation to see whether these linguistically-motivated compound terms are actually of value for retrieval.

Example 2

This illustrates information seeking, taking Vickery et al's work on a front end prototype to show AI used for a specific subfunction. This is a system designed to elicit a sufficiently rich problem statement and to convert it to a search query, as originally tested for gardening literature. The system uses a lexicon (supplemented by a classification scheme) to supply frames for request words. These frames are of different types and have slots also of different types, which can be filled by other frames, so we can think of the slots as representing the different facets of a topic. The aim is to fill the slots using any available information or by asking the user further questions, operating in a rule-based way. The system then converts the filled frame structure for the problem statement into a Boolean term query.

This is thus an intermediate example, since while it is primarily an application of AI-derived rule-base technology, it also exploits some world knowledge in its frame set and some, albeit very minimal, inference. But it is not really a self-contained subsystem, since it is tied to larger document retrieval operations.

Example 3

This example illustrates special-purpose processing of an AI kind is it might be done as one complete component of a larger multi-purpose information system. It is thus contrasted, as a complete embedded subsystem, with the previous example

where both the material involved, and the use of it, are much less restricted.

The example is Young and Hayes' message processing system. This is designed to interpret banking telexes, converting untidy natural language texts into standard-form records in a database (from which summaries may also be produced). In essence the system makes an initial categorisation of messages by key words, and then uses a set of syntactic-semantic pattern associated with the selected category to analyse the individual message sentences, basically as predicate-argument structures, but for the particular domain of banking. The system can then build an actual record as an instantiation of the category frame, by using rules to assign the various items given by the sentence phrase analysis to appropriate slots and to make inferences to obtain missing information.

In all of these example systems, the use of AI is feasible because the demands it makes are offset by limitations elsewhere. In the first case, indexing is a limited task, so full natural language understanding is not needed. In the second, the information seeking task is less limited, but the system is operating within a specialised domain and can exploit the notion of normal topic types for requests within it. In the third the task is less limited still, but is applied to very specialised material for which regular conceptual content can be assumed.

At the same time these systems, though they are limited in various ways, all depend on substantial resources, like grammars and rule sets, devised by human beings; and indeed in all such systems there has to be a humanly provided knowledge base, however much it may be extended with particular data derived from its processing inputs or even, ideally, with general information through learning. So there are issues about the return to be expected from the investment in providing the base, and also about the general IT resources, like interface capabilities, systems like these may assume. There are clearly cases where limited and especially embedded AI-derived systems and functions may be really valuable, for instance a rule-based system for standards conversion or one for automatic descriptor assignment. But it is a crucial empirical fact about these systems that it is impossible to significantly extend their scope.

Final comment

In evaluating the potential for AI, it is essential to be aware of how little we can do. This becomes only too obvious if we take just the simple case represented by natural language interaction with conventional, coded databases, and consider what the backend system has to do to handle users' questions effectively. Natural language access to databases is fine (or good enough) when everyone knows what sort of thing is in the database and is rather careful about framing questions. But suppose we have the standard 'Suppliers and Parts' example, where parts are described by type (e.g. nut, screw), weight, and colour, and the user asks "Does Smith supply brass screws?" If the system response is "No" this is liable to be completely misunderstood, but if the system cannot interpret "brass" because it cannot recognise that it is not dealing with screw weight or colour, it cannot help the user by pointing out the error in the question. However to provide the helpful response

that the database does not cover the materials of which parts are made, the system needs a classification structure indicating the range of intrinsic properties that parts can have, including properties the database in fact does not cover.

Unfortunately, even doing this will not suffice. If the user asks "Does Smith supply wood screws?", this refers to an extrinsic part property, function, and not necessarily to the database's default assumed function for all screws. The system will thus fail in the same way as before, through not having a sufficiently rich encompassing domain model, if it does not alternatively fail because it does allow for (other) parts which may be made of wood and assumes screws may be made of wood too. How is the system to indicate what is wrong with the user's question, or rather how are we to provide for its being able to do this?

References

Sparck Jones, K, and Tait, J.I. 'Automatic search term variant generation', Journal of Documentation 40, 1984, 50-66.

Vickery, A. et al 'A reference and referral system using expert system techniques', Journal of Documentation 43, 1987, 1-23.

Young, S.R. and Hayes, P.J. 'Automatic classification and summarisation of banking telexes', Proceedings of the Second Conference on Artificial Intelligence Applications, 1985, 402-408.