THE GENERATION OF TERM DEFINITIONS FROM AN ON-LINE TERMINOLOGICAL THESAURUS

John McNaught

Centre for Computational Linguistics
UMIST
P.O. Box 88
Manchester UK

ABSTRACT

A new type of machine dictionary is described, which uses terminological relations to build up a semantic network representing the terms of a particular subject field, through interaction with the user. These relations are then used to dynamically generate outline definitions of terms in on-line query mode. The definitions produced are precise, consistent and informative, and allow the user to situate a query term in the local conceptual environment. The simple definitions based on terminological relations are supplemented by information contained in facets and modifiers, which allow the user to capture different views of the data.

I Introduction

This paper describes an on-going project* being carried out at UMIST, which is concerned with the nature, construction and use of specialised machine dictionaries, concentrating on one particular type, the terminological thesaurus (Sager, 1981). The system described here is capable of dynamically producing outline definitions of technical terms, and it is this feature which distinguishes it from other automated dictionaries.

II Background

A. Published Specialised Dictionaries

These traditional reference tools, while often containing high quality terminology collected after painstaking research, do not in the normal case afford the user an overall conceptual view of the subject field, as they exhibit a relative lack of structure. Moreover, due to the limitations of the printed page, the format of the entries is fixed, such that users with differing information needs are obliged to search through to them irrelevant data. The only real aid which allows the user to place a term roughly in the local conceptual environment is the conventional definition. However, such definitions tend to be idiosyncratic, inconsistent and non-rigorous, especially if the subject field is of any great size. Contexts, while of some help, are notoriously difficult to find and control, and should only be seen as supplementary to a rigorous definition of the term which firmly places the term in the conceptual space. Those reference tools containing definitions which are rigorous exist mainly in the form of glossaries established by standards bodies. However, standardisation of terminologies is a slow affair, and is restricted to certain key terms or fields, such that no overall conceptual structure may be obtained from these glossaries.

B. Term Banks and Machine Dictionaries

Term Banks offer many advantages over traditional dictionaries, and are becoming more and more common, especially among organisations which have urgent terminology needs, such as the Commission of the European Communities or the electronics firm Siemens AG in W. Germany. National bodies likewise use term banks to control the creation and dissemination of new standardised terms, e.g. AFNOR (France) and DIN (W. Germany). In the UK, work is going ahead, coordinated by UMIST and the British Library, to set up a British Term Bank. Other important term banks exist in Denmark (DANTERM) and Sweden (TERMDOK). However, despite this growth in the number of term banks and other computer based dictionaries, there remains a sad lack of overall structuring of the terminological data. In some cases, dictionaries have been transferred directly onto computer, in other cases, data base management considerations have overriden any attempt at systematic terminological representation of the data. Some term banks have made provision for expressing relations between terms (AFNOR, DANTERM) but these relations are not as yet exploited to their full.

C. Documentation Thesauri (DTs)

These tools, whether on-line or published from magnetic tape, represent gross groupings of terms (via descriptors) for the purpose of indexing and retrieval of documents. A hierarchical structure is apparent in a thesaurus, with general relationships being established between descriptors, such as BT (Broad Term), NT (Narrow Term) and RT (Related Term). Some thesauri further distinguish e.g. BTG (Broad Term Generic), NTP (Narrow Term Partitive) and so on. However, by its very nature and purpose, a DT is merely a tool for selecting and differentiating between the chosen items of the artificial reference system of

* sponsored by the Department of Education and Science through the award of a Research Fellowship in Information Science

an indexing language. The existence of overlapping and even parallel indexing languages attests the inadequacy of DTs for representing generally accepted terminological relationships. Other problems associated with DTs are highlighted when attempts are made to merge DTs and to match descriptors across language boundaries. Existing DTs also find great difficulty in representing polyhierarchies (Wall, 1980) hence the ambiguous nature of the RT relation. The best known attempt at solving such difficulties is the THESAUROFACET (Aitchison, 1970).

D. Terminological Thesauri (TTs)

Traditionally, the TT (as advocated by e.g. Wüster, 1971) represents relationships between concepts rather than descriptors in as much detail as possible. As such, it has mainly been the preserve of terminologists. The TT has the advantage of precisely situating a term in the conceptual environment, through making appeal to relationships such as generic and partitive (and their various detailed subdivisions), and to relations of synonymy (quasi-, full synonyms, etc) and antonymy. A classic example of the TT approach to structuring data is the Dictionary of the Machine Tool (Wüster, 1968), which has served as a basis for the present project.

However, although systematic in conception and detailed in execution, this particular work displays the constraints inherent in the Wüsterian approach, which is akin to that of the DT, namely reliance on the hierarchy as a structuring tool. For example, given the partial sub-tree in figure 1a.:

```
PRINTER
[...]
 PAPER TRANSPORT MECHANISM
  FORM FEED
   FEED RATE CONTROL
    TAPE CONTROLLED
     TAPE CONTROLLED PRINTER
```

figure 1a. Problems with a hierarchy

we would like to be able to relate TAPE CONTROLLED PRINTER to its true superordinate, PRINTER, to say that it is a type of printer. Again, given the structure in figure 1b.:

```
CHARACTER
[...]
 PRINTABLE
  PRINT CHARACTER
  CONTROL CHARACTER
```

figure 1b. Problems with a hierarchy

we would like to be able to represent the relationship of CONTROL CHARACTER to CHARACTER directly. This is impossible in the hierarchical approach, where one is constrained to adopt one scheme, and to represent only one possible relationship, whereas a term may have multiple relationships to multiple terms. As with DTs then, conventional TTs are incapable of representing one-to-many and many-to-one relationships.

E. Summary

There exists a need for a representational device which can capture the necessary relationships between terms in a natural and informative manner, and which is not constrained by the limitations of the printed page, or the mental capacity of the terminologist.

III The on-line Terminological Thesaurus

The present project has concentrated on finding a device capable of responding to the demands of different users of terminology, and which would allow a systematic representation of terminological data. We have retained the term Terminological Thesaurus, but have given it a new meaning. The particular device we have constructed combines the advantages of the conventional TT (systematic structure, relationships) and of the traditional dictionary (definitions). This is achieved by using inter-term relationships first to construct a highly complex network of terms, and subsequently, at the retrieval stage, to generate natural language defining sentences which relate the retrieved term to others in its terminological field. This is done by means of templates, such that the user is presented with an outline definition of a term (or several definitions, if a term contracts relations with more than one term) which will help him to circumscribe the meaning of the term precisely. Although the particular orientation of the project is to generate definitions, the semantic network that is constructed could be used for other ends, and future work will investigate these possibilities. We stress here that the definitions that are produced are not distinct texts stored in the machine and associated with individual terms; rather, the declared relationships between terms are used to dynamically build up a definition, and terms from the immediate conceptual environment are slotted into natural language defining templates. These definitions have the advantages of being precise, system internal and always correct, providing the correct relationships have been entered. Preliminary work in this area was first carried out at UMIST in the late 70s, when the feasibility of using terminological relationships to structure data was shown, and an experimental system was implemented, based on a hierarchical representation, that output simple definitions (Hann, 1978). This was found to be inadequate, for the reasons outlined above, hence the adoption in the present project of a richer data structure.

The data base for the system is then a semantic network. As with most semantic networks, the most one can really say about it is that it consists of nodes and arcs: terms form the nodes, and relations between terms the arcs. In actual fact, the data base consists of several files, with the character strings of terms being assigned to one file, such that all search and creation operations for the network proper are carried out

using simply logical pointers to bare nodes carrying the geometrical information needed to sustain the network, thus avoiding the overhead of storing variable length strings often in duplicate. A virtual memory has been implemented such that file accesses are kept to a minimum, and all pointer chains are followed in fast core. The basic data structure of the network is the ring, and the appearance of the network is that of a multiway extended tree structure. Facilities exist for on-line interactive creation and search of the network. An important design principle is that the computer should relieve the terminologist (or indeed the naive user) of the burden of keeping track of the spread and growth of a conceptual structure. We have already seen how the hierarchical approach to terminology failed to account for all the facts, and forced the terminologist into misrepresenting or distorting the conceptual framework. With a network, ease and naturalness of representation is achieved, but at the cost of increased complexity for the human mind. Thus a human will quickly lose track of the ramifications of a network, even if he could represent these adequately on some two dimensional medium. Entrusting the management of the network to the computer ensures precision and consistency in a very large data base.

At the input stage, in the simple case, the terminologist need give only 3 pieces of information: two terms and the relationship between them. As the system is open-ended by design, the terminologist can declare new relationships to the system as he works, i.e. it is not necessary to firstly elaborate a set of relationships. Further, neither of the two input terms need necessarily be present in the data base. If both are absent, the system will create a closed sub-network, which will only be linked to the main network when other links are made with one or both of these terms. As input proceeds, one may have the (perhaps non-consecutive) inputs

<X rel A> <Y rel A> <Z rel A>

where {X,Y,Z,A} are terms and <rel> a relationship. The system will link all terms related to <A> in a ring having <A> flagged as the 'head' node. Thus the terminologist is not required to overtly state the relationships between {X,Y,Z}. Leaving the computer to establsih links among terms from an initial single input relationship ensures high recall. Note that choice of refined relationships aids high precision, although too many refinements may be detrimental to retrieval, in which case some automatic mechanism for widening the search to include closely associated relationships would be necessary. However, this would imply that information be conveyed to the system regarding the associations between relationships, and would be a strong argument in favour of designing a set of relationships prior to the input of terms. At present, we have no strong views on this subject. The system is open-ended to accept new relationships; it is up to the terminologist how he organises his work.

In the complex case, where there are perhaps several terms having the same relationship as the input term to a common 'head', or where the 'head' may have several sub-groups (q.v.) associated with it, the system interacts with the user to tell him there are several possibilites for placing a term in the network, and shows him structured groups of brother terms having the same relationship as the input term to the 'head', where his input term may fit in. It is important to realise that the user need have no knowledge of the organisation of the network. He is asked to make terminological decisions about how an input term relates to others in the immediate conceptual environment.

The notion of 'sub-group' is the only one which requires explanation in terms of the theory behind the organisation of the system. This notion was introduced in an attempt to represent the fact that there may be terms that are mutually exclusive alternatives, and which attract other terms which can cooccur without restriction. A simplistic example will make this point clearer. For the sake of discussion, we assume the following parts of a radio, shown in figure 2.:

RADIO
VALVE
TRANSISTOR
AERIAL

figure 2. Simplified parts of a radio

what we wish to represent is the fact that if a radio has valves, it has no transistors, and vice versa, but whichever is the case, there is always an aerial present. What has happened here, terminologically, is that there are two terms missing from the concept space, referring to the concepts 'valve radio' and 'transistor radio' respectively. Or it may be the case that the terminologist has not as yet entered the generic subdivisions of radio. Thus there are two 'holes' here, as yet unfilled by a term. The solution adopted, is to create dummy nodes in the network, which act as ring 'heads' for sub-groups each of which contains one of the mutually exclusive alternatives, plus any terms that are strongly bound to one or both of the alternatives, but not themselves mutually exclusive. The dummies refer back directly to the true head term, and may be converted at any time into full nodes if the terminologist's answers to questions about his input indicates that a new term ought to occupy this position, with this particular relation to the original head term and with this particular sub-group of terms. Terms which are common to all sub-groups, and which have a relationship to the original head term, are merely inserted in the ring dominated by the original head, and are by default interpreted as belonging to all sub-groups. In our present example, this would apply to 'aerial'. Various checks are incorporated to prevent e.g. terms common to all sub-groups being bound to all these groups - that is, if one binds a term to every possible sub-group under an original head, this would imply that it does not in fact have any special binding power, or cooccur only with terms in these sub-groups. The resulting

structure for this admittedly simple example is shown in figure 3, where primed nodes are dummy nodes dominating a sub-group ring.:

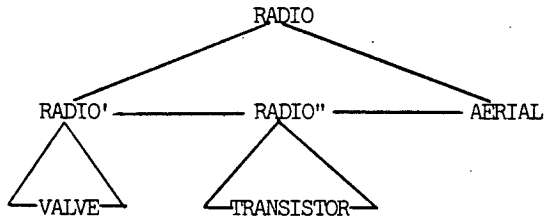

figure 3. Representation of alternatives

The basic data structure, with terms ontologically related to another term being logically subordinated to it, and with several other relations being established either automatically or semi-automatically in response to user interactions, provides enough information for the generation at search time of outline definitions of terms. The main file containing the semantic network proper has the record structure shown in figure 4.:

Field	Value	Type
1	RELATION	CHAR
2	MODIFIER	CHAR
3	FACET	INTEGER
4	FATHER/BROTHER	INTEGER
5	SON	INTEGER
6	VARIANT	INTEGER
7	CONTENT	INTEGER
8	ALTERNATIVE	INTEGER
9	FLAGS	INTEGER

figure 4. Network file record structure

The FLAGS field apart, all integer fields are logical pointers to other records in the network file, except for CONTENT which points into another file containing records which give information on the actual character strings of terms. Most of the field values are self-explanatory. The FATHER/BROTHER field has a dual value (indicated by an appropriate flag) and togetner with the SON field is used to build the basic ring structure.

The VARIANT field is used to form another ring which links nodes representing the same term in relation to different 'heads', and is commonly employed to represent polyhierarchies, which as will be recalled posed a problem for DTs and TTs. Here the advantage of the CONTENT pointer becomes apparent, as only the geometrical network-sustaining information is duplicated when a term enters into relation with more than one 'head'.

Two fields remain which require more detailed explanation, namely the MODIFIER field and the FACET field. These were introduced to enhance the outline definitions the system produced, which, although precise and consistent, were found to be rather uninformative in certain respects. For example, to generate the definition 'A vernier is a type of scale' leaves something to be desired, when the definition in Wüster's dictionary refers to 'a small movable auxiliary scale'. One could of course get round this by declaring a new type of scale to the system, namely 'auxiliary scale' or even 'movable auxiliary scale', if this were terminologically acceptable. We think though that to append 'small' would be stretching things rather far. However the introduction of a MODIFIER field allows some measure of finer description, by allowing the user to specify an adjective or adjectival phrase, which in this case, and perhaps commonly, would be relational, i.e. 'vernier' is seen as small in relation to a larger 'scale', but may be large with respect to e.g. 'microvernier'. The modifier is thus attached to the geometrical, relational node of the network, not to the content, string bearing node.

The FACET field takes its name from the facets well-known in the construction of DTs. A facet is here used in a similar manner to a DT facet, that is, as a classificatory tool, to give a different view of the data. A facet represents a gross grouping of terms according to some feature. Examples of facets are:

BY DIRECTION BY METHOD OF PROPULSION
BY MATERIAL BY APPLICATION
BY SHAPE BY METHOD OF OPERATION,
etc.

In traditional DT work, though, a descriptor can appear only under one facet. In the present system, a term can appear under many facets. This gives extreme flexibility and allows the terminologist to draw fine and not-so-fine distinctions between groupings of terms. In most DTs, there is little attempt at structuring facets - they are used in a fairly ad-hoc manner. In the context of the present project, research is being carried out by Catherine Yarker into the nature of facets, which will shed light on how they could best be employed in the system. An interesting point to note is that what are normally called terminological relationships could justifiably be viewed as a subset of facets, the difference being that they are more commonly used, display more structure, and have undergone systematic investigation over the years.

Output from the system is available in a variety of formats, depending on how much, or which type of, information the user desires. There now follow a few examples which show the potential of the system:

Query: CAR
Response: CAR is a type of VEHICLE, together with BUS, LORRY, TRAIN and TRACTOR.

Q.: PYLON
R.: PYLON is a part of WINDMILL, together with VANE and GENERATOR.

These show how a simple definition of a term is given, by relating it to its generic or partitive superordinate, and listing other terms having the same relationship to the superordinate as the query term.

Q.: SLUICE VALVE
R.: SLUICE VALVE is a part of PUMP
<more?>
Yes
R.: SLUICE VALVE is a type of GATE VALVE

Q.: GATE VALVE
R.: GATE VALVE is a type of VALVE together with PISTON SLIDE VALVE, PLUG VALVE, LIFT TYPE VALVE and CHECK VALVE

Here the user asks for a further definition of SLUICE VALVE, then queries GATE VALVE.

The addition of facet information, as we have pointed out, gives rise to more informative definitions, in that the user can see how a term is differentiated from other terms in the immediate environment, for example:

Q.: FLANGE
R.: FLANGE is a component of FLANGED FITTING, and is related to STEEL FLANGE by material, to DETACHABLE FLANGE by property, to CONNECTING FLANGE by method of connection, to BOSSED FLANGE by shape and to FLAT FACED FLANGE by flange face configuration

Q.: STEEL FLANGE
R.: STEEL FLANGE is a type of FLANGE distinguished by material. Other types of FLANGE are DETACHABLE FLANGE and LOOSE FLANGE distinguished by property, CONNECTING FLANGE, SCREWED FLANGE and WELDED FLANGE distinguished by method of connection, BOSSED FLANGE and OVAL FLANGE distinguished by shape and FLAT FACED FLANGE, RAISED FACE FLANGE and FULL FACED FLANGE distinguished by flange face configuration.

Experiments are still under way to determine how best to use facets, and how best to formulate the definitions. It appears useful, in a definition, first to relate a term to another by a common terminological relationship (part of, type of) and then to refine the definition by bringing in facets.

There is also the possibility to ask for a specific relationship, for example, if one were to ask for parts of a wheel, the display might read:

WHEEL is composed of HUB, SPOKE, RIM, WHEEL CENTRE and TYRE.

The usefulness of more refined terminological relationships is shown by the following examples:

KEY is a part of KEYBOARD
WHEEL is a part of CAR
RADIO is a part of CAR
ENGINE is a part of CAR

where the standard 'part of' relationship proves inadequate. Therefore, we introduce subdivisions of the partitive relationship, which generate the following outputs:

KEY is an atomic part of KEYBOARD (i.e. the latter consists wholly of the former).
One or several WHEELs are contained in CAR
RADIO is an optional part of CAR
ENGINE is a constituent part of CAR (i.e. CAR contains other parts, including ENGINE)

These few examples hopefully give some indication of the system's potential. With a complex network enriched with refined terminological relationships, modifiers and facets, we can look forward to the generation of extended, informative definitions. It may be argued that problems could arise in maintaining the consistency of the network, however the interactive input procedure is designed to show the consequences of a particular choice or insertion before the input is recorded definitively in the network. Nevertheless, there comes a point when one has to rely on the user himself not to make silly decisions. Due to the extreme flexibility of the system, and the use of a network as a representational device, the terminologist is free to introduce whichever relationships he desires, and to link whichever terms he chooses. This freedom may be anathema to those who adhere to the rigorous hierarchical approach to terminology, however, used with judicious care, the system is capable of recording multiple relationships in a way denied to the proponents of the hierarchical approach, which in the end provide a basis for the generation of information that is more fully developed, and more illuminating due its richness.

In the near future, an interactive editor will be implemented to help the terminologist adjust the data base, in case of error, or to monitor the changes brought about by the a change of relationship, facet, etc.

It should be noted that the system is designed to be multilingual, and is capable of outputting foreign language equivalents. As we have chosen to deal with rather normalised terminology, we make no claims as to the capability of the system to handle more general vocabulary, where there would be sometimes radical differences between the conceptual systems of different languages. At the moment, we work purely with one-to-one mappings across language boundaries. However, unlike the traditional term bank, which merely enumerates foreign language equivalents, this system, on the other hand, upon addressing a foreign language equivalent in the data base allows immediate entry to a ring of foreign language synonyms, from which the entire parallel conceptual network of the foreign terminology may be accessed. The possibility is then open for further definitions in the foreign language to be output, if desired.

IV IMPLEMENTATION

The system is completely written in 'C', a general purpose system programming language, and is implemented on a Z-80 based S-100 microcomputer, with 64kbyte memory and a 33mbyte hard disk. When the system is eventually stable, a virtual memory routine written in assembly language by Sandra Waites will replace the existing 'C' routine, to speed up access times. The system runs to several thousand lines of code, including utilities and

basic input/output functions ('C' provides none of the latter) and is split into several chained programs, for reasons of memory space restrictions. Execution time is not therefore as fast as it could be, although the hard disk does make a substantial difference to access times. When mounted on a 16-bit microcomputer running under the Unix operating system, as is envisaged in the near future, and equipped with improved index searching routines (not a primary purpose of the project), there should be little delay in response time.

For reasons of economy and experimentation, the basic network file record is limited to 16 bytes (see figure 4 above), however, in a future version of the system, other features may be added, for example a ring head pointer in each record, to save scanning all ring records to the right of the entry point to find the head. Further, the content file record, which contains information on character strings, could be expanded to hold the types of information found in traditional term bank records, e.g. grammatical class, context, author, date of entry, sources, etc. This would then imply that a full-blown term bank could be set up, organised around a semantic network, such that the bank would be structured according to terminological criteria, not to data base management criteria.

V ACKNOWLEDGEMENTS

I would like to thank Sandra Waites and Catherine Yarker for their valuable contribution towards the realisation of this system, and my colleagues Rod Johnson and Professor Juan Sager for their advice during the course of the project.

VI REFERENCES

Aitchison, J. The Thesaurofacet: A Multi-Purpose Retrieval Language Tool. J. Doc., 1970, 26, 187-203.

Hann, M.L. The Application of Computers to the Production of Systematic, Multilingual Specialised Dictionaries and the Accessing of Semantic Information Systems. Manchester, UK : CCL/UMIST report, 1978.

Sager, J.C. Terminological Thesaurus. Lebende Sprachen, 1982, 1, 6-7.

Wall, R.A. Intelligent indexing and retrieval: a man-machine partnership. Inf. Proc. & Man., 1980, 16, 73-90.

Wüster, E. The Machine Tool : an Interlingual Dictionary. London, UK : Technical Press, 1968.

Wüster, E. Begriffs- und Themaklassification. Nachrichtung für Dokumentation, 1971, 22:4.