The Identification of Important Concepts in Highly Structured Technical Papers

Chris D. Paice and Paul A. Jones

Department of Computing
Lancaster University
Bailrigg
Lancaster LA1 4YR
U.K.

Abstract

Automatic abstracting, typically based on extraction of important sentences from a text, has been treated as a largely separate task from automatic indexing. This paper describes an approach in which the indexing and abstracting tasks are effectively combined. It is applicable to highly structured empirical research papers, whose content can be organised using a semantic frame. During a scan of a source text, stylistic clues and constructs are used for extracting candidate fillers for the various slots in the frame. Subsequently, an actual concept name is chosen for each slot by comparing the various candidates and their weights.

Sentence Extraction

In 1958 H.P. Luhn published an influential paper entitled "The Automatic Creation of Literature Abstracts." In it he proposed that abstracts might be generated automatically by selecting from a source text sentences which contained strong clusters of 'significant words' [Luhn 1958]. Each potential cluster would receive a score reflecting the number of significant and non-significant words in it, and each sentence would receive the score of the highest-scoring cluster in it, if any. Those sentences whose scores exceeded some set threshold would be extracted for inclusion in the abstract.

In order for this approach to work, Luhn had to address the question of how to recognise the significant words. These were seen to be content words which were particularly associated with the subject matter of the document in question. They were recognised by first using a table to eliminate prepositions, articles, conjunctions and other common function words, then accepting as significant the most frequent of the remaining word-stems.

ACM-SIGIR'93-6/93/Pittsburgh, PA, USA

Luhn's paper served as a setting-off point for subsequent research into both automatic abstracting and statistical indexing. Considering that abstracting and indexing are both concerned with identifying and expressing the central content of a document, it is striking that since Luhn's paper research in these two areas has remained on two almost entirely separate tracks.

To a large extent, the reason for this is that it was quickly realised that the presence of key-word clusters is by no means the only clue to sentence significance; indeed, research by Edmundson seemed to indicate that it was a rather poor clue [Edmundson 1969]. As a result, researchers tended to concentrate their efforts on the other clues, including:

- the *position* of a sentence within the document or paragraph [Baxendale 1958; Edmundson 1969];
- the presence of *cue words* and expressions such as "important", "definitely", "in particular" (all positive), and "unclear", "perhaps", "for example" (all negative) [Edmundson 1969; Rush *et al.* 1971];
- the presence of *indicator constructs* such as "The purpose of this research is" and "Our investigation has shown that" [Paice 1981];
- the number of semantic links between a sentence and its neighbours [Skorokhod'ko 1972].

The above abstracting methods are often referred to as *extraction procedures*. The only serious attempt at a working system of this kind is represented by the Chemical Abstracts Service's *ADAM* system, which relied mainly on the use of an extensive list of cue words, most of them negative ones [Pollock & Zamora 1975].

Improved extraction of the most appropriate sentences from a source text would seem to be a matter of incremental tuning. Unfortunately, however, a collection of extracted sentences often shows a marked lack of coherence with, most glaringly, the frequent presence of 'dangling anaphors'. *Anaphors* are words such as pronouns, demonstratives and comparatives which can only be understood by referring to an *antecedent* appearing

earlier (or occasionally later) in the text. If an extracted sentence contains an anaphor but no antecedent then it is at best jarring, and at worst unintelligible.

Unfortunately, handling these cases automatically entails a host of problems. For example, we need to be able to determine *(a)* whether a potentially anaphoric word is actually being used in an anaphoric sense or not; *(b)* whether an anaphor has an antecedent within the same sentence ('internal') or elsewhere ('external'); and *(c)* whether an external anaphor has a local antecedent (typically in the previous sentence) or a remote one. Remote antecedents are usually referenced by noun phrases starting with "the", such as "the lava flow". In such a case the antecedent may match the anaphoric phrase exactly ("a lava flow"), inexactly ("lava was flowing"), or implicitly ("an erupting volcano").

Despite a good deal of study, it cannot be said that the dangling anaphor problem has been satisfactorily solved [Paice & Husk 1987; Paice 1990]. This being so, we cannot expect the sentence extraction method to produce a good standard of output, at least in the near future.

Text Summarisation

An alternative approach, making use of techniques from artificial intelligence, entails performing a detailed semantic analysis of a source text, and so constructing a semantic representation of the meaning of the document. A set of frames, tailored to the domain of application, is normally used to facilitate the analysis and representation tasks. When analysis is complete, output templates are used to generate a textual summary from the instantiated frames [DeJong 1982; Rau *et al.* 1989].

Unfortunately, the knowledge base required for a system of this kind is of necessity large and complicated, and is moreover specific to the domain of application. Some idea of the complexities involved can be gained from the descriptions by Hahn of his TOPIC system [Hahn 1989, 1990]. Although the performance of such systems may be reasonable for documents of a narrow domain and specific genre, there seems little prospect of broadening such systems to cope with a much wider variety of input.

Genre and Domain

Although by contrast the sentence extraction approach may appear to be indifferent to the particular nature of the source text, this is really just a difference of degree. First, the extraction methods are designed to process expository technical texts, such as empirical research papers, and are unlikely to work well on (say) political news reports. Secondly, there is reason to believe that an extraction system will work best if tailored to the intended subject domain; thus, the ADAM system was deliberately tailored for abstracting chemistry papers [Pollock & Zamora 1975]. Hence, although reasonable flexibility is obviously desirable, complete source independence is not to be expected. As regards genre, we may assume that the input to a typical abstracting program will consist of technical papers or articles. As regards domain, we can accept a certain amount of tailoring provided that *(a)* output of some kind will be produced even for an unfamiliar domain, and *(b)* adapting the system for a different preferred domain will not be forbiddingly difficult.

A deficiency of the basic extraction approach is that it does not take any systematic account of the structure of the source document. It is well known that any normal technical paper is organised in quite a predictable way [van Dijk 1980; Kircz 1991; Paice 1991]. Moreover, abstracts of technical papers are also structured [Liddy 1991], and there appears to be a strong resemblance between the typical structures of both source texts and abstracts. This has led Paice to suggest the use of a so-called *abstract-frame* to guide the extraction of sentences from a source document, thus improving the balance and coverage of the resulting abstract [Paice 1990].

Examination of a selection of empirical research papers quickly reveals that the stereotyping is not just a matter of overall organisation, but is also stylistic and semantic. By 'semantic' we mean that the main concepts discussed in these papers fit into a narrow and largely predictable range of roles. By 'stylistic' we mean that the message of such papers is conveyed by a variety of characteristic constructs and expressions. Our hypothesis is that these characteristic constructs provide evidence about the semantic roles of the concepts bound to them. Thus if we find the construct "the effect of *X* on *Y*", it is reasonable to assume that *X* denotes an independent variable or influence, and *Y* a dependent variable or property. This kind of stylistic regularity gives us a means for identifying the main concepts discussed in a document.

The proposed method resembles the AI-type text summarisation approach in that it revolves around the identification of concepts which are inserted into a kind of frame. However, there is no attempt to produce a representation of the full meaning of a document. Moreover, the use of superficial stylistic features means that concept identification is relatively fast and simple. The inherent unreliability of stylistic clues is offset by the fact, remarked long ago by Luhn, that important concepts tend to be mentioned many times in a paper [Luhn 1958]. If many of these mentions are in distinctive contexts, then a list of several candidate strings may be compiled for a given conceptual role; it then remains to pick out the 'best' name from among these candidates.

One attraction of the proposed method is that, by focusing on concept identification rather than sentence extraction, it offers the prospect of a system in which indexing and abstract generation may be performed as interrelated activities.

The Concept Identification System

At the present stage, our research has focused mainly on papers in the field of crop agriculture, and in Figure 1 we list some of the most obvious semantic roles for such papers. Our present set of stylistic context rules provides

for the first nine of these; the remainder are some examples of future possibilities.

Our semantic categories were constructed during manual analysis of a selection of research papers in the area of crop agriculture. Though developed independently, our categories somewhat resemble those used in PLEXUS, an expert system for answering questions about gardening [Vickery & Brooks 1987].

In general, we look to find positive evidence for the various concepts in a paper. Some conceptual roles may simply not be applicable to a particular document, while in other cases there may be a natural default value. Thus, in Figure 1, 'high-level properties' are those broad properties of crops which are presumed to be sensitive to pests, fertilizers, management methods or other influences. The usual high-level property of a crop is the *yield*, and this is assumed by default unless some other property is indicated. Possible alternatives include growth rate and germination rate.

The reader may notice that none of the headings in Figure 1 refers to the *results* of experiments. It is clear that extraction of significant findings from research papers is a harder task than identification of the properties listed. This, in part at least, is because the headings in Figure 1 refer to individual concepts, or lists of individual concepts, whereas the findings of an investigation are typically expressed as correlations between pairs (or sometimes larger *n*-tuples) or properties. Thus, full conceptual analysis of the Results and Discussion sections of a paper is left as a problem for future consideration. Obviously, since the system described here relies on analysis of running text, no attempt is made to analyse actual tables or graphs of results.

This restriction means in effect that our method produces abstracts which are indicative but not informative in content; that is, they provide a statement of the topic but no indication of the outcome of the research. In an effort to overcome this deficiency, our program when possible chooses one or more whole sentences from the final paragraphs of a paper, as described later; in other words, some informative material is added using an old-fashioned extraction method.

Selection of Filler Strings

Our program makes use of a collection of context patterns defining the various stylistic constructs and the conceptual roles which are associated with them. The source text of a paper is scanned looking for instances of any of the patterns; each time a pattern is found a 'filler string' is extracted from the associated context and added to a list of candidate names for the relevant conceptual role.

The context patterns were selected by manual analysis of the corpus. For each concept a list of sentences where the concept occurred and its position in the text was made. From these lists common patterns were identified. In many cases a particular pattern included more than one concept, as can be seen in figure 2. Figure 2 includes some examples of context patterns, and in this representation possible alternatives are shown, both for literal words within patterns (e.g., "of/in") and for conceptual roles (e.g., "SPECIES/AGENT"). The pattern format used by our existing program is more primitive and repetitious than this, but is logically equivalent.

Semantic Roles		Explanation
SPECIES	:	The crop species concerned
CULTIVAR	:	The cultivars (varieties) used
HIGH-LEVEL PROPERTY	:	The property being investigated (e.g. yield, growth rate)
PEST	:	Any pest which infests the crop
AGENT	:	Chemical or biological agent applied
INFLUENCE	:	e.g. drought, cold, grazing, cultivation system
LOCALITY	:	where the study was performed
TIME	:	years when the study was conducted
SOIL	:	description of soil
CLIMATE	:	climate at LOCALITY
TREATMENT	:	e.g. management techniques (more specific than INFLUENCE)
LOW-LEVEL PROPERTIES	:	properties which are measured (e.g., dry weight, size of roots)
PROCESS	:	e.g. uptake, assimilation
NUTRIENT	:	e.g. Potassium, Nitrogen (K, PO_3)

Semantic roles for crop agriculture papers.

figure1.

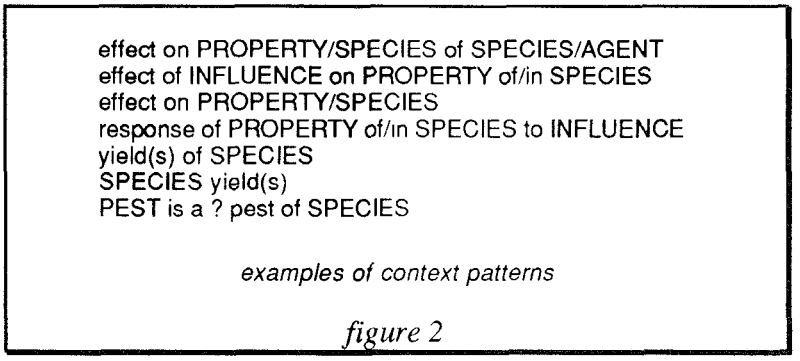

figure 2

The query symbol in Figure 2 shows a position where a number of intervening words may occur in the source text; thus the pattern

"PEST is a ? pest of SPECIES"

would match the sentence

"A.lolii is a common pest of ryegrass"

with "A.lolii" being a candidate for PEST, "common" being ignored as it coincides with "?" and "ryegrass" being a candidate for SPECIES.

It is clear that some of the constructs in Figure 2 provide much stronger evidence than others, and accordingly each construct has an associated weight (not shown in Figure 2). At present the weights are integers ranging from 1 to 10, with 10 denoting very strong evidence and 1 or 2 indicating a bare possibility.

The weights were initially assigned by pure guesswork, but they and the context patterns have been systematically adjusted during successive trials with a development corpus. It is hoped that a scheme may later be devised to perform this automatically via a feedback system.

The content of acceptable filler strings is subject to certain restrictions. Two situations may be distiguished: *(a)* the conceptual role occurs at the right or left hand end of a pattern, as in "yield of SPECIES"; *(b)* the conceptual role appears inside a pattern, as in "effect of INFLUENCE on...."

In case *(a)* the filler string has to be terminated at some point on the right or left, as appropriate. This is achieved by maintaining a list of 'forbidden words', containing commonly occurring verbs together with most prepositions and other function words. A candidate filler string is thus picked up word by word, moving right or left as appropriate, until either a forbidden word or a sentence boundary is encountered.

In case *(b)*, the candidate filler string is obviously delimited by the construct which contains it. However, in this case the filler strings are subjected to a 'trimming' procedure. This removes any 'forbidden' words from both ends of the filler string, thus ensuring that the two cases generate clues of the same style.

The reader may object that these crude provisions cannot be relied on to produce a sensible and well-formed concept name. This is not usually too serious, since the actual choice of concept name is performed at the next stage.

Choice of Concept Name

When the above stage is complete, each relevant conceptual role should possess a list of candidate filler strings each with an associated weight. For each role which has more than one candidate, the program must identify a string to serve as a suitable concept name for output purposes.

The present procedure is very simple, and consists of selecting from among the candidates that substring (or more properly, that whole-word sequence) which commands the highest aggregate weight. Thus, a substring which occurs in three or four of the candidate contexts should be favoured over a word or phrase which appears only once or twice. In cases where there are no repeated substrings, the highest-weighted filler string is used in its entirety. If the highest aggregate weight is less than 4 no concept name is selected, and the concept will be ignored on output.

Figure 3 illustrates this procedure for the SPECIES concept of one of our agriculture documents. Note that the processed text always includes the title of the paper, since this is recognised as a particularly fruitful source of information.

Generation of Abstracts

In order to generate the final abstracts, an output template is used into which the names chosen for the conceptual roles are inserted. Obviously, some of the less important roles may not be instantiated, and the form of the output is then adjusted accordingly. When fully developed, the output templates will provide for alternating forms of expression, in order to avoid the output looking too fixed and repetitive.

Title : The effect of mildew seed treatment and foliar sprays used alone or in combination in 'early' and 'late' sown Golden Promise spring barley, Aberdeen, 1976 to 1982.

Sentences from document :

Powdery mildew has consistently been the most damaging foliar disease of spring barley in Britain causing reduction in grain yield of around 10 per cent in England and Wales.

In Scotland the problem has been accentuated by the widespread growing of the highly mildew susceptable cultivar Golden Promise.

In Scotland however during this period Golden Promise on average occupied over 50% of the total spring barley seed area certified with a low of 41% in 1977 and a peak of 60% in 1981.

The following series of experiments was conducted against this background of the predominance of a highly susceptable spring barley cultivar grown under conditions increasingly more conducive to early mildew development.

Rule which matched	Weight	Candidate string
From Title :		
effect of ? in ? sown SPECIES	9	Golden Promise spring barley, Aberdeen, 1976
effect of ? in SPECIES	7	combination
in ? in SPECIES	4	'early' and 'late' sown Golden Promise spring barley, Aberdeen, 1976
? of SPECIES	2	mildew seed treatment and foliar sprays
From text :		
disease of SPECIES	4	spring barley
? SPECIES cultivar	4	the highly mildew susceptible
? by SPECIES	2	the widespread growing
? with SPECIES	2	a low
? of SPECIES	2	the total spring barley seed area certified
? SPECIES cultivar	4	a highly susceptable spring barley

About ten other weak clues (i.e. 'of SPECIES', 'with SPECIES') were identified, but they are not shown here as they did not affect the result and would unnecessarily complicate the example.

The following lists the substrings which occurred more than once.

Substring	Number of Occurrences	Weight	
spring barley	5	9+4+4+2+4	=23
Golden Promise spring barley, Aberdeen 1976	2	9+4	=13
highly	2	4+4	=8
susceptible	2	4+4	=8
mildew	2	2+4	=6
seed	2	2+2	=4

Therefore 'spring barley', with a total weight of 23, is selected as the filler for the concept SPECIES.

Example of selection of the filler string for the semantic role SPECIES

figure 3

The generation program has already been described in another paper, and will not be described in detail here [Jones & Paice 1992]. Figure 4 shows the output obtained for four of our test papers; abstract A to C appear satisfactory, but abstract B contains obvious errors.

We earlier mentioned the difficulty of extracting informative material, especially information about the findings from a piece of work. In an attempt to mitigate this, our program incorporates tests for a number of *indicator constructs*, such as "results indicate that...", "we have shown that..." and "in conclusion". When these are found the complete sentence is appended to the abstract.

Some of the examples in Figure 4 include such sentences, with the indicator constructs shown in italics. Unfortunately, although these sentences are often very helpful, they introduce the danger, mentioned earlier, that external anaphoric references may be introduced.

Abstract A
Title :- The assesment of the tolerance of partially resistant potato clones to damage by the potato cyst nematode Globodera pallida at different sites and in different years.
Citation :- Ann. Appl. Biol., 1988, 113, pp79-88

This paper studies the effect the pest G. pallida has on the yield of potato. An experiment in 1985 and 1986 at York, Lincoln and Peterbourgh, England was undertaken. These *results indicate* clearly *that* there are consistent differences between potato cultivars in their tolerance of damage by PCN as measured by proportional yield loss.

Abstract B
Title :- The relationship between leaf canopy development and yield of barley.
Citation :- Ann. of Appl. Biol., 1988, 113(2), pp 357-374

This paper studies the effect of leaf canopy development on the yield of barley. An experiment in 1979-80 and 1980-81 at Edingburgh was undertaken on a sandy loam overlying clay loam classified soil, using the cultivars Golden Promise (GP) and Marris Mink (MM) and the winter cultivars Video (V) and Marris Otter (MO). *The approach described in this paper* could be used as a basis for describing and modeling the growth of barley particularly with regard to comparisons between cultivar od agronomic treatments.

Abstract C
Title :- Soil temperature and water content, seeding depth and simulated rainfall effects on winter wheat emergence.
Citation :- Agronomy Journal, 1989, 81, pp 609-614

This paper studies the effect of soil temperature on the emergence of winter wheat. The cultivar 'Norstar' hard red winter wheat was used. *The results provide* insight into the western Canadian farmers' dilemma of whether to direct-seed (stubble-in) winter wheat into a dry seed bed at the optimum date or wait for some precipitation before sowing.

Abstract D
Title :- The effect on winter wheat of grazing by Brent Geese *Branta Bernicla*
Citation :- Journal of Applied Ecology, 1990, 27, pp 821-833

This paper studies the effect of Brent Geese Branta on the each field a grid of winter wheat. The experiment took place at Deepdale Marsh, Burnham, Deepdale. The fact that ear density increased due to grazing in one field *indicates that* there is probably little value in the farmer sowing seed at a higher density in an attempt to compensate for geese grazing.

Automatically generated abstracts

figure 4

Evaluation

Evaluating the quality of abstracts has always seemed an ill-defined task involving a large measure of subjective judgment. Indeed, the question as to whether an abstract is 'good' or 'bad' depends most critically on the requirement of the person reading the abstract. For instance, if a user investigating the growth of potatoes on peat soils is presented with Abstract A in figure 4, they may consider this to be a 'bad' abstract as it does not mention the type of soil used in the trial. However, if at a different time the same user was investigating the effect of pests on potatoes they would consider this a 'good' abstract as it includes the detail they are interested in.

The obvious method to evaluate abstracts is by conducting a user trial with a known corpus and a set of questions (similar in style to the standard collections held for IR evaluations). Unfortunatly the corpus of papers we have is not large enough to conduct this style of trial and the lack of papers in machine readable form means that expanding this corpus is a very time consuming task.

A number of alternative methods seem to be feasible. The first is to use 'expert' abstractors from an abstracting company. This would involve asking them to identify any weaknesses or omissions in the automatic abstacts. Unfortunately the abstractors at CAB International are used to producing *informative* and not *indicative* abstracts, and we therefore feel that any results they produced would be biased against the style of abstracts we are attempting to generate.

Evaluation could involve comparison of each automatic abstract with a specially constructed *target abstract.* However, the notion of there being a single 'ideal' abstract by which to judge an automatic abstract is obviously

flawed and in practice would still involve subjective decisions being made by the evaluator.

In an attempt to involve possible users in the process of evaluation a small trial was organised. In this trial we provided each user with a full paper and asked them to select the important parts of it. We were hoping that these lists of important sections could them be compared to the abstracts to see how many of the selected points were covered. However, the users tended to select sections very heavily biased towards their own particular interest and in order to acquire the style of output we required the users had to be directed to such an extent that any results obtained would tend to be biased in favour of our work.

Due to these problems we have decided to present the evaluation of the system as a group of statistics. The statistics indicate which concepts were and were not included in the generated abstracts. This should provide the reader with enough information for them to be able to assess the effectivness of this system, without having to resort to ambiguous terms such as 'good' or 'bad'.

The statistics are split into three sections, *focal concepts*, *non-focal concepts* and *conclusions*. The focal concepts are those considered vital to the paper (i.e. for Abstract A: G. Pallida, potato and tolerance are focal, whereas for Abstract B: leaf canopy development, barley and yield are focal).

The non-focal concepts are all the other relevant concepts for a paper which may be identified by the system, and the conclusions are the summarising sentences discussed above. These three types are shown in tables 1, 2 and 3 respectively.

The table heading 'Too Long' refers to final candidate phrases which contain the correct concept, but also contain extra words. The table heading 'Too Short' refers to clues which have been correctly identified but which have been over-shortened by the final candidate selection process (e.g., "barley" where "spring barley" would be more appropriate).

The 'original rules' referred to are those developed just on the original corpus. The 'modified rules' are those developed on the original corpus and the previously unseen corpus together.

The last entry in each table is labelled 'Animal husbandry'. This small test was performed as a first step to evaluating the domain (in)dependance of this particular set of context patterns. No changes were made to any parts of the system before this trial was conducted. As the trial was so small (6 documents) no conclusions can be drawn from it, but we were pleasantly surprised at the system's success at identifying focal concepts and conclusions in these papers.

Table 1: Focal concepts

Development corpus (24 documents), original rules.

	Correct	Incorrect	Missing	Too Long	Too Short
Number in Corpus	55	0	9	2	4
Percentage	79%	0%	13%	2%	6%

Unseen corpus (24 documents), original rules

	Correct	Incorrect	Missing	Too Long	Too Short
Number in Corpus	46	14	4	3	6
Percentage	64%	19%	6%	3%	8%

Development corpus, modified rules

	Correct	Incorrect	Missing	Too Long	Too Short
Number in Corpus	59	4	0	1	5
Percentage	86%	6%	0%	1%	7%

Unseen corpus, modified rules

	Correct	Incorrect	Missing	Too Long	Too Short
Number in Corpus	59	5	2	1	5
Percentage	82%	7%	3%	1%	7%

Animal husbandry papers (6 documents), modified rules

	Correct	Incorrect	Missing	Too Long	Too Short
Number in Corpus	12	3	0	0	3
Percentage	67%	17%	0%	0%	16%

Table 2: Non-focal concepts

Development corpus, original rules.

	Correct	Incorrect	Missing	Too Long	Too Short
Number in Corpus	43	7	24	3	2
Percentage	55%	9%	30%	4%	2%

Unseen corpus, original rules

	Correct	Incorrect	Missing	Too Long	Too Short
Number in Corpus	17	7	37	2	2
Percentage	26%	11%	57%	3%	3%

Development corpus, modified rules

	Correct	Incorrect	Missing	Too Long	Too Short
Number in Corpus	40	4	31	3	1
Percentage	51%	5%	39%	4%	1%

Unseen corpus, modified rules

	Correct	Incorrect	Missing	Too Long	Too Short
Number in Corpus	25	2	32	3	3
Percentage	38%	3%	32%	5%	5%

Animal husbandry papers, modified rules.
Only 2 non-focal points were selected from the papers (both correct): one was a location the other was the timing of the study.

Table 3: Conclusion sentences

Development corpus, original rules.

	Correct	Incorrect	Missing	Anaphor present
Number in Corpus	15	0	6	2
Percentage	65%	0%	26%	9%

Unseen corpus, original rules

	Correct	Incorrect	Missing	Anaphor Present
Number in Corpus	5	1	14	0
Percentage	25%	5%	70%	0%

Development corpus, modified rules

	Correct	Incorrect	Missing	Anaphor present
Number in Corpus	16	0	5	2
Percentage	70%	0%	22%	8%

Unseen corpus, modified rules

	Correct	Incorrect	Missing	Anaphor Present
Number in Corpus	8	3	7	2
Percentage	40%	15%	35%	10%

Animal husbandry papers, modified rules

	Correct	Incorrect	Missing	Anaphor Present
Number in Corpus	4	0	2	0
Percentage	67%	0%	33%	0%

Discussion

It is noticable that as the rules were refined an improvement occurred in the focal point hit rate (see Table 1, original rules cf. modified rules), whereas the non-focal hit rate decreased (Table 2). We believe that this is due, at least in part, to the relatively large number of occurrences of each focal concept. This meant that we had a larger corpus with which to define the focal rules. In many cases non-focal concepts are not easily identifiable as the lack of examples has made it harder to identify the common patterns which appear around them. If a larger corpus were developed we believe that that the hit rates for these non-focal concepts could be increased.

Identification of focal concepts is also aided by the fact that these concepts occur in typical positions within the text and usually have more than one occurrence. The non-focal concepts do not seem to appear in such fixed locations and usually only occur once. Having to identify just this one context pattern is far more problematic than being able to identify any one of a number of the contexts in which the focal concepts occur. We consider it unlikely that, using context patterns alone, identification of non-focal concepts will ever reach the 80%-plus success rate achieved for the focal concepts.

The success rate for conclusion sentences increased as more rules were added and the weights refined (Table 3). It seems likely that the focal point and conclusion sentence rules are fairly well optimised and any other alterations will only cause small changes which may be either positive or negative. However the non-focal context patterns certainly need further work.

Future Plans

The next stage in this work will be the implementation of an automatic weight adjustment/pattern highlighter system. It is hoped by using this tool on a large corpus that the non-focal hit rate can be improved and that the focal and conclusion sentence hit rates can be shown to be valid.

Other work which is required includes improvement of the concept selection method. At present where there is little or no duplication among the candidate strings for a concept, there is an obviously likelihood of an overloaded name being assigned. Contrarily, incomplete names can easily be produced by the present rather crude name choosing method; thus, if three candidate strings contain "spring barley", and another contains "barley" alone, then the total of four occurrences of "barley" will cause that to be chosen as the concept name. Some of these problems may be overcome by fuller linguistic processing of the candidate strings.

Another method which we hope to explore is to incorporate a domain thesaurus into the system; the listing of a term such as "spring barley" in the thesaurus would cause an increase in the weight recorded for that string. It is likely that use of a thesaurus would also improve the identification of some of the non-focal concepts, such as soil type.

Attention needs to be given to the handling of coordinated noun phrases, to the scientific names of species, and to the handling of dates.

At a later stage, we will need to measure the retrieval performance resulting from using our concept selection tool as a source of index terms. However, this will require the processing of large numbers of source documents and so cannot be completed quickly.

Acknowledgements: we wish to express our thanks to CAB International of Wallingford, Oxfordshire, for providing financial support to one author, and for their advice and encouragement throughout this project.

REFERENCES

Baxendale, P.B. 1958: "Man-made index for technical literature an experiment", *IBM Journal of Research & Development* **2**(4), 354-361.

DeJong, G. 1982: "An overview of the FRUMP system", in W.G.Lehnert & M.H.Ringle (eds.), *Strategies for Natural Language Processing*, Hillsdale, N.J.: Lawrence Erlbaum Associates.

Edmundson, H.P. 1969: "New methods in automatic extracting", *Journal of the Association for Computing Machinery* **16**(2), 264-285.

Hahn, U. 1989: "Making understanders out of parsers: semantically driven parsing as a key concept for realistic text understanding applications", *International Journal of Intelligent Systems*, **4**(3), 345-385.

Hahn, U. 1990: "Topic parsing: accounting for text macro structures in full-text analysis",*Information Processing & Management* **26**(1), 135-170.

Jones, P.A. & Paice, C.D. 1992: "A 'select-and-generate' approach to automatic abstracting", 14th British Computer Society Information Retrieval Colloquium, Lancaster, March 1992.

Kircz, J.G. 1991: "The rhetorical structure of scientific articles: the case for argumentational analysis in information retrieval", *Journal of Documentation* **47**(4), 354-372.

Liddy, E.D. 1991: "The discourse-level structure of empirical abstracts: an exploratory study", *Information Processing & Management* **27**(1), 55-81.

Luhn, H.P. 1958: "The automatic creation of literature abstracts", *IBM Journal of Research & Development* **2**(2), 159-165.

Paice, C.D. 1981: "The automatic generation of literature abstracts: an approach based on self-indicating phrases", in R.N.Oddy *et al* (eds), *Information Retrieval Research,* London: Butterworths.

Paice, C.D. & Husk, G.D. 1987: "Towards the automatic recognition of anaphoric features in English text: the impersonal pronoun *it*", *Computer Speech & Language* **2**, 109-132.

Paice, C.D. 1990: "Constructing literature abstracts by computer: techniques and prospects",*Information Processing & Management* **26**(1), 171-186.

Paice, C.D. 1991: "The rhetorical structure of expository texts", Proceedings of the Informatics 11 conference, York, March 20-22, 1991.

Pollock, J.J. & Zamora, A. 1975: "Automatic abstracting research at the Chemical Abstracts Service", *Journal of Chemical Information & Computer Science* **15**(4), 226-232.

Rau, L.F., Jacobs, P.S. & Zernik, U. 1989: "Information extraction and text summarization using linguistic knowledge acquisition", *Information Processing & Management* **25**(4), 419-428.

Rush, J.E., Salvador, R. & Zamora, A. 1971: "Automatic abstracting and indexing. II. Production of indicative abstracts by application of contextual inference and syntactic coherence criteria", *Journal of the American Society for Information Science* **22**(3), 260-274.

Skorokhod'ko, E.F. 1972: "Adaptive method of automatic abstracting and indexing", *Information Processing 71*, North Holland; pp.1179-1182.

van Dijk, T.A. 1980: *Macrostructures*, Hillsdale, N.J.: Lawrence Erlbaum Associates; chapter 3.

Vickery, A. & Brooks, H.M. 1987: "PLEXUS - the expert system for referral", *Information Processing & Management* **23**(2), 99-117.