Devika V. Aptagiri, M.A. Gopinath, A.R.D. Prasad
Documentation and Research and Training Centre, Bangalore, India

A Frame-Based Knowledge Representation Paradigm for Automating POPSI

Aptagiri, D.V., Gopinath, M.A., Prasad, A.R.D.: A **frame-based knowledge representation paradigm for automating POPSI**
Knowl.Org. 22(1995)No.3/4, p.162-167, 11 refs.
This paper is based on the project work carried out by the authors at Documentation Research and Training Centre. Knowledge representation models are used in building intelligent systems for problem solving. The paper discusses, a frame based knowledge representation model built for automatic indexing. The system assigns POPSI indicators and produces subject strings for titles. The results are given in appendices.

(Authors)

1. Introduction

An Index is designed as a tool for information retrieval. To achieve maximum efficiency an index should cater to the different approaches of the users. The usual approaches are by author, title and subject. Subject indexing deals with the formulation of subject strings to satisfy the subject approach of the users. This process involves analysis of the document for its subject content and modulation of the indexing string. The earlier attempt in subject indexing was UNITERM indexing. As subject terminology comprises compound terms, the uniterm index does not fully justify its role in information retrieval especially not in a manual environment (although post-coordination of terms is much easier in a machine environment). Compound terms are better dealt with in pre-coordinate indexing. However, compound terms alone without the contextual information may tend to decrease precision in information retrieval.

One of the first attempts in providing context in automatic indexing is the KWIC (Key Word In Context) index, which has introduced the concept of permuting the constituents of entries in a subject index. However, as the KWIC index permutes the titles of documents there is no guarantee that the entries carry the contextual information sufficiently if not fully. This lacuna becomes even more apparent in the case of fancy titles which give no clue of the subject of the document. In other words, the success of KWIC and its variants like KWOC, KWAC etc. depends on the expressiveness of titles and fails when titles are not expressive.

The next move in context based subject indexing was the emergence of PRECIS (PREserved Context Indexing System), which uses the permutation technique of KWIC, but unlike KWIC, the subject strings are based on an analysis of the thought content of a document. POstulate based Permuted Subject Indexing, POPSI, propounded by G. Bhattacharyya is a context based indexing system. POPSI has incorporated the principles of classification according to the Ranganathan School of Thought, in order to give context to terms in subject strings . The present work deals with the automation of POPSI.

Devika V. Aptagiri (b. 1967) Project Assistant at the DRTC. Research undertaken: Design of a knowledge representation model for the management of an analytico synthetic classification system. Mysore University. Areas of Interest: Artificial Intelligence, Automatic IRS, Multimedia, Internet. Ongoing Work: Project on Application of OCR in building bibliographic databases. 7 publications

M.A.Gopinath (b.1944) Professor and Head, DRTC. ADIS from DRTC. PhD from Karnatak University, Dharwad, India. 30 yrs of teaching Areas of Interest: Classification, Indexing, Expert Systems. Over 250 publications, Editor of the journal Library Science with a Slant to Documentation and Information Studies. Author of several books.

A.R.D.Prasad (b.1954) MA, Mphil in philosophy, ADIS form DRTC. PhD on 'Application of Natural Language Processing (NLP)techniques to Indexing Languages' from Karnatak University, Dharwad, India. Lecturer at the DRTC. Specialisation in application of Artificial Intelligence to Library and Information Science. Areas of Interest: Artificial Intelligence, Expert System, NLP, Multimedia, Internet. Over 25 publications.

2. POPSI

Postulate-Based Permuted Subject Indexing (POPSI) developed by G. Bhattacharya uses the Analytico-Synthetic method for string formulation and permutation of the constituent terms in order to satisfy different approach points to the document. It is guided by accepted postulates and principles.

2.1 Constituents of POPSI

2.1.1 Lead Heading

This is the heading of the index entry. The index term which appears in the position of a lead heading is called a 'Lead Term'. The lead term decides the position of that

entry in the entire subject index and thereby ensures precision in search. The lead term is also referred to as 'Access Term' or 'Approach Term'. The lead heading may also contain other terms with or without auxiliary/ function words that further qualify the lead term, making its meaning clear.

Lead terms are generally presented in noun form. They may also consist of auxiliary/function word (symbols) as and when warranted. The auxiliary symbols serve as connectors to constitute words of a lead term, so that the lead term becomes readable and conveys a specific meaning. An advantage of connectors is that they resolve ambiguity in case of certain noun phrases in natural language. An altemative to using the connectors is to have a predefined order in presenting the components of a lead term.

2.1.2 Context Heading

A context heading generally appears as a next line to the lead heading. A context heading may contain a few more subject terms along with auxiliary words. The purpose of a context heading is to provide the information about the context in which the lead term occurred.

2.2 Subject Indexing Languages (SIL)

2.2.1 Structure

According to the SIL Theory different structures with regard to subject indexing language statements are identified as Semantic Structure, Elementary Structure, and Syntactic Structure

2.2.2 Semantic Structure

The semantic structure of a subject, essentially refers to its parts, species and different concepts and their relation with each other. In other words, the name of a subject comprehends all its divisions and subdivisions and what they denote. This is essentially due to the expression of relations between different concepts. For example, Philosophy as a subject comprehends ethics, epistemology, metaphysics etc. The semantic structure is based on 'Genus-Species', 'Whole-part', 'Broader subject/ narrower subject' or 'Extension' or range of the subject.

2.2.3 Elementary Structure

The components of a subject index statement may belong to more than one elementary category depending on the semantic significance of each category. The structure recognized on the basis of elementary categories to which the different components of a subject index statement belongs to the 'Elementary Structure'. This structure is artificially postulated. For example, S.R. Ranganathan has identified a set of fundamental categories viz. 'Personality', 'Matter', 'Property', 'Energy', 'Space', 'Time'. This categorization is similar to that of parts-of speech in natural languages, or phrases in phrase structure grammars of linguists. The syntax of a natural language is defined in terms of parts of speech, whereas the syntax of a subject indexing language is defined in terms of predefined categories.

2.2.4 Syntactic Structure

The syntax of a language prescribes rules for valid or acceptable sentences. In a way, it expresses the relation of elementary categories to each other, in a given language. The syntax of subject indexing languages also sets rules for the construction of valid subject indexing statements.

In natural languages, for a given set of elementary categories, the syntax may allow a number of valid sentences to be formed, resulting sometimes in ambiguity. However, in the subject index languages the syntax should not allow ambiguous statements. As SILs are artificial languages, they are supposed to have well-defined rules for the construction of valid statements. The syntax of Subject Indexing Languages should clearly state the order of its elementary categories.

The concept of deep structure, has attracted many researchers in the field of Subject Indexing. If the same advantages are promised in the case of subject indexing languages, it can well be concluded that a subject indexing statement of a given subject indexing language can well be translated into a statement of another subject indexing language.

To achieve this, G. Bhattacharyya has attempted to study various subject indexing languages in order to identify the deep structure of SIL. The Phrase Structure Grammar identifies some basic categories of words like Noun phrase, Verb Phrase, Noun, Verb, Determiner etc. In the case of subject indexing languages Bhattacharyya arrived at elementary categories, viz. Discipline, Entity, Property and Action (Acronym DEPA), and a special component called 'Modifier'.

Discipline (D)
Discipline is an elementary category that includes conventional fields of study or any aggregate of such fields or artificially created analogous fields. For example: Library Science, Artificial Intelligence, Computer Science, Radiology etc.

Entity (E)
Entity includes manifestations of having perceptual correlates, or only conceptual existence, and distinct from the properties and actions performed by them or on them. For example: Library Collection , Lung, Patent, etc.

Property (P)
Property includes manifestations denoting the concept of attribute-qualitative or quantitative. For example: Efficiency, Specific Gravity, Precision, Disease.

Action (A)
Action includes manifestations denoting the concept of doing. Action may manifest as self action or as external action. For example: Function, Migration are self actions; Selection, Evaluation are external actions.

Modifier (M)
In relation to the manifestation of any one of the elementary categories, 'Modifier' refers to qualify the manifesta-

tions without disturbing its conceptual wholeness. For Example: 'Subject' in Subject Classification.
A modifier can modify a manifestation of any of the fundamental categories as well as a combination of two or more manifestations of two or more elementary categories. Modifiers generally create speeies/types. Modifiers can be either Common modifiers like Form, Time, Environment and Place or Special modifiers which are Discipline-based, Property-based or Action-based.

2.3 The Syntax of the Deep Structure (DS) of a Subject Indexing Language

The basic rule of syntax associated with the DS of a SIL is that 'Discipline' should be followed by 'Entity' (either modified or unmodified) appropriately interpolated or extrapolated wherever warranted by 'Property' and/or 'Action' (both modified or unmodified).

A manifestation of Property follows immediately the manifestation in relation to which it is a property. A manifestation of Action follows immediately the manifestation in relation to which it is an action. Property and Action can have another property and / or Action directly related. In other words, the rules of syntax relating to Property and Action of another property or action is that their positions must always follow the Property or Action to which they are related.

The Rules of Syntax relating to species / Part and Modifiers according to POPSI are:

A species or part follows immediately the manifestation in relation to which it is a species or part. A modifier follows immediately the manifestation in relation to which it is a modifier. This rule also applies to cases where there are more than one modifier to the same manifestation. If more than one sequence of modifiers to the same modifyee is equally valid (in terms of its representation in natural language), according to the above rule, the choice of any sequence is acceptable (3).

In general, the rules of syntax give rise to the following syntactical structure.

DISCIPLINE followed by ENTITY followed by PROPERTY and /or ACTION. PROPERTY and /or ACTION may further be followed by PROPERTY and /or ACTION as the case may be. Each of the above components may further admit of, and be followed immediately by their respective SPECIES/TYPES and/or PARTS and/or SPECIAL MODIFIERS. The COMMON MODIFIERS generally occur last in the sequence. These rules of syntax are in total conformity with Ranganathan's theory of Classification, (especially with the Principle of facet sequence) based on the principle of decreased concreteness, (wall picture principle) and its derivations like Actand-Action-Actor-Tool principle.

3. An AI Model for Subject Indexing

To solve complex problems in artificial intelligence there is a need for a large amount of knowledge and some mechanisms for manipulating that knowledge to create solutions to new problems. The different ways of representing knowledge show that specific knowledge representation models allow for more specific and powerful inference mechanisms that operate on them.

The three major models of knowledge representation can be enlisted as
1) Rule-based models, 2) Semantic Nets, and 3) Frame-based models

3.1 Rule-based Models

Rule-based knowledge representation depends on a number of rules and works deductively. A set of if/then rules are incorporated and identification of facts/truths is done by definite prestated combinations of such rules.

3.2 Semantic Nets

Semantic Nets are knowledge representation models where information is represented as a set of nodes connected to each other by a set of labeled arcs, which represent relationships among nodes. These nets are represented internally using some kind of attribute-value memory structure.

3.3 Frame-Based Knowledge Representation

A frame is a data structure for representing a stereotyped situation. It is basically a network of nodes and relations. The top levels of a frame are fixed and represent things that are always true about the supposed situation. The lower levels have many terminal-slots that must be filled by specific instances of data.

Two different kinds of entities dealt with in knowledge representation are:

(a) Facts: Things that we want to represent
(b) Formalism: Representations of facts in a chosen formalism.

The present work adopts Frame Based Knowledge Representation for representing the facts in the system. Inheritance in frames is used in assigning categories to the constituents terms of the subject heading.

4. Objectives of the Project

The objectives are a) to build an Intelligent System for information retrieval based on POPSI, b) to incorporate Artificial Intelligence techniques in the design of the system, aand c) to demonstrate the viability of Intelligent Indexing systems.

4.1 Hypotheses

A. Knowledge is expressed in Natural language and knowledge representation is nothing but semantic representation.
B. A frame-based knowledge representation model is ideal for semantic representation.
C. The Analytico-Synthetic approach is amenable to frame-based knowledge representation.
D. Every noun phrase has a definite role to play in a subject string that is used as a surrogate to a document.
E. The modulated subject string provides context to each individual key term in a subject string and this modulation could be

achieved by using a hierarchical representation of key term.
F. Each individual word knowledge representation can be treated as an object. In an object oriented approach to knowledge representation, the main task is to identify the objects, their properties and the relation among objects to each other.

4.2 System

The present work aims at building a frame-based representation model to generate the subject strings in D E P A (POPSI) syntax. The program is written in PROLOG. The steps through which the system was developed are as follows:

a) Compilation of a system knowledge base (Appendix 1)
b) Design of an Inference Engine (Appendix 2)
c) Sample of Output Subject Strings (Appendix 3)

The input for this are noun phrases from selected titles in Library Science. The noun phrases collected are passed on to a knowledge based system in order to represent the semantics. The knowledge based system comprises a frame-based knowledge representation, modeled on the lines of subject classification; to be specific: with a deep structure of the indexing language as represented by the DEPA sequence.

Actions, operations and processes are stated as nouns rather than verbs, e.g. design, application, production, acquisition, execution, elimination, implementation, validation, substitution, determination, measurement, synchronization etc; or are stated as activity nouns derived form verbs like: programming, editing, processing/ instructing, indexing, estimating, coding, etc. The basic parsing task in query analysis is the recognition of noun phrases and the relation between them (10).

Representation of categories is attempted by building a frame based representation for semantics. Therefore, the basic approach is to build a Frame Based Knowledge Representation model, where each term (noun phrase) is treated as an object. The primary task is to express each object (noun phrase), its properties, and its relation to others objects.

4.3 Frame Formulation

The objective in interpretation of a sentence is to get a set of interrelated frames, in which each frame represents a verb or noun and its associated modifiers. Frames contain slots that can be unified or filled. Filled slots represent facts.

Unifying means finding substitutions of terms for variables to make logical formulae identical. A matching operation is carried out between frames to fill in the value of the frame necessary for representation. For example,

```
frame-1 : Cataloguing
              category : action
frame-2 : Cataloguing
              category : X
```

This goal frame can be interpreted as the query 'To which category 'Cataloguing' belongs?' Matching it against the fact frame above yields the substitution

X = action

which is the answer to the query.

5. Knowledge Representation of Elementary Categories

5.1 Discipline

The natural language processing system developed here is domain specific. The subject index entries belong to the discipline 'Library Science'. The frame-based knowledge representation model intended to represent the semantics of subject index entries belong to the discipline 'Library Science'. Accordingly the system fills in the discipline slot, the value 'Library Science'. The predicate fillin' in the program adds the necessary list structure [1, Library Science] where '1' is the indicator value of the discipline category.

5.2 Entity

The category entity includes all kinds of libraries. The relation among different isolate ideas in this category or that of part-of relation. For example

```
value(acquisition section , part_of, library).
valuc(library,category, entity).
```

The input strings for the system are taken in the area of Library Science. These are essentially noun phrases taken after parsing selected titles.

5.3 Property

The property category includes all belongings of libraries like library collection, library rules etc. The relation between different types of documents is represented by the kind_of relation. The facts in property look like

```
valuc (Reference books, kind_of, books).
value (books, kind_of,library collection).
value(library collection, part-of,library).
value(library collection, category, property).
value(property, indicator, [ ' 6 ', ' .2 ' ] ) .
```

The system gathers all the information for the noun phrase such as, what are its super-ordinates, to what category it belongs, and what is the value of the indicator. For example, if the noun phrase Reference books' is encountered, further information is filled in the property slot,

(6.2 library collection 6.2 books 6.2 reference books)

However, if the document does not explicitly deal with a particular kind of document the system produces the following string to be filled in the property slot.

(6.2 library collection)

5.4 Action

The system emphasizes this category specially when compared to entity or property. The other categories are dealt in order to give context to this category.

```
value (administration, use, management).
value (administration, category, action).
value (action, indicator, 6.2.9).
```

If the syntactic parser produces a noun phrase (management^techniques) then, the system produces the following string to be filled in the action slot.

(6.2.9 administration)

5.5 Modifiers to Action

The modifiers mostly belong to the modifiers of kind 1 category, which form complex terms like 'management using linear programming'. For which the system produces

(6.2.9 management 6.2.9.5 (using) linear programming)

from the following facts in frame representation.

value (management, category, action).
value (linear^programming , category,
modifier_to_action).
value (modifier_to_action, indicator, 6.2.9.5).

However, a problem arises when the expressive title contain more than one isolate belonging to the modifier category. However, the program factors such a composite term (where the components belong to more than one category). Here, it should be emphasized that the order of quasi isolates decides the order of the elements.

6. Formulation of Subject Entries

The system adopts the following steps in order to arrive at the POPSI subject strings.

Step 1: Preparation of expressive title

This step is done manually. The expressive titles thus prepared are input to the system for the generation of subject strings. The titles have been input in list structures as shown below:

[a, practical, manual, of, colon, classification]
[administration, of, technical, libraries]
[a, physical, bibliography, for, librarians]

Step 2: Factoring the expressive titles for constituent terms.

This step involves the construction of noun-phrases which include uniterms, compound and composite terms. It is hypothesized that every noun phrase has a definite role to play in a subject string that is used as a surrogate to a document. For example:

Title: 'A Practical manual of Colon Classification'

Represented as: [a, practical, manual, of, colon,classification]
Noun phrases generated: practical^manual
colon^classification

Step 3: Standardisation of terms

Standardisation of non-standard terms is done by incorporating the information in facts in representation. Reference is given from a non-standard term to a standard term. The system picks up the standard term and then assigns the category and the indicator accordingly. For example:

Title:	'administration of technical libraries'
Representation:	[administration, of, technical,libraries]
	(administration, use,management).

The system picks up the term 'management' for the subject strings in place of the term 'administration'.

Step 4: Modulation

In this step the system adds on the additional information such as the basic subject and then arranges the constituent terms according to the DEPA syntax prescribed by POPSI. Since the titles chosen for input are from the field of Library Science the system fills the data as follows

Discipline = Library Science

Further it assigns each of the constituent term to its category and then arranges the terms by either extrapolating or intrapolating for broader or narrower terms in the subject string. For example:

Title:	[administration, of, libraries]
Noun Phrases:	administration library
Representation:	(administration, use, management).
	(management,category, action).
	(library, category, entity).
Modulated String:	Library Science, library, management

Step 5: Assignment of indicators

POPSI assigns indicators to each constituent term to facilitate the sequence of arrangement of terms in the subject string. The system assigns the prestated indicators to the constituent of the modulated string. For example:

Title:	[administration, of, libraries]
Noun Phrases:	administration library
Representation:	(administration, use, management).
	(management,category, action).
	(library, category, entity).
	(action, indicator, ['6', '.1']).
	(entity, indicator,['6']).

From the above facts the systems assigns the indicators to the constituent terms as shown below:

6 library
6.1 management

Step 6: Generation of subject entries using the syntax as prescribed by POPSI.

Finally the subject string is formulated in which the constituent terms are assigned indicators and the output is consistent with the syntax DEPA of POPSI. For example:

Title:	[administration, of, librarics]
Noun Phrases:	administration library
Representation:	(administration, use, management).
	(management,category, action).
	(library, category, entity).
	(action,indicator, ['6','.1']).
	(entity, indicator, ['6']).

Index string: Library Science, 6 library, 6.1 management

7. Conclusion

The system tackles the different steps in indexing after the formulation of the expressive title which is the only step done manually. Standardisation of non-standard terms to standard terms is taken care of by the system. The

frame-based knowledge representation system assigns the indicators according to the categories DEPA prescribed by POPSI and automatically generates the subject strings as given in the appendix 3.

References

(1) APTAGIRI, D. V.: Design and Development of a Knowledge Representation Model for Automatic Indexing. Guided Research Project. Documentation Research and Training Centre. Bangalore. 1994.

(2) BHATTACHARYA, G.: A General theory of Subject Indexing Languages. PhD Thesis. Karnatak University. Dharwad. 1980.

(3) BHATTACHARYA, G.: Subject Indexing language: Its theory and Practice. In: DRTC Refresher Seminar Volume. DRTC, Bangalore. Oct. 1981.

(4) BORKO, H., BERNIER, C.: Indexing Concepts and Methods. New York: Academic Press 1978.

(5) GAZDAR, G., MELLISH, C.: Natural Language Processing in Prolog.: An introduction to computational linguistics. New York: Addison-Wesley 1989.

(6) GOPINATH, M. A., PRASAD, A. R. D.: A Knowledge Representation Model for Analytico-Synthetic Classification. Advances in Knowledge Organisation. Vol 4. (1994) p.320-327.

(7) RAY, M., SPARKJONES, K.: Automated Language Processing. In: Ann. Rev. Inform. Sci. & Technol. vol.6. ed.by C.A. Cuadra. Chicago 1971. p.141-66.

(8) PRASAD, A. R. D.: Application of Natural Language Processing Tools and Techniques in Developing Subject Indexing Languages. PhD. Thesis. Dharwad: Karnatak University 1994.

(9) SCHANK, R.: The Cognitive Computer: On language learning and Artificial Intelligence. Addison-Wesley 1984.

(10) VICKERY, B., VICKERY, A.: Application of Language processing for Search Interface. J.Doc. 48(1992)No.3, p.255-75

(11) WINSTON, P.H.: Artificial Intelligence. Ed 2. London: Addison-Wesley 1984.

APPENDIX 1

THE SYSTEM KNOWLEDGE BASE

```
% Indicators
value (time,            indicator,   ['1']).
value (space,           indicator,   ['3']).
value (property,        indicator,   ['.2']).
value (entity,          indicator,   ['6']).
value(action,           indicator,   ['.1']).

% Space
value (world,  category,    entity).
value (asia,   bt,          world).
value (europe, bt,          world).
value (india,  bt,          world).

% Entity
% Types of libraries :
value(library,              category,   entity).
value(archive,              usc,        library).
value(information^centre,   usc,        library).
value(referral^centre,      usc,        library).

% Sections of a library
value(library^divisions,    category,   entity).
value(circulation^section,  bt          library^divisions).
value(periodical^section,   bt,         library^divisions).
value(acquisition^section,  bt,         library^divisions).

%Types of documents
value (document,            category,   entity).
value (newspaper,           bt,         document).
value (standard,            bt,         document).
value (patent,              bt,         document).

% Action
% Technical Processing
value (technical ^processing,  category,   action).
value (document^acquisition,   bt, technical^processing).
value (cataloguing,            bt, technical^processing).
value (classification,         bt, technical^processing).

% Information Service
value(information^serrvice,    category,   action).
value(documentation^service,   usc, information^service).
value(reference^service,       bt,  information^service).
```

APPENDIX 2

SAMPLES OF SUBJECT STRINGS GENERATED

Title: A history of library associations.
[a,history,of,library,associations]
history
library^association
Library Science. 6 learned^body professional^association

Title: A physical bibliography for librarians.
[a,physical,bibliography for librarians]
physical^bibliography
librarian
Library Science 6 document bibliography.
physical^bibliography

Title: A practical manual for Colon Classification.
[a,practical,manual,for,colon,classification]
practical ^manual
colon ^classification
Library Science 6.1 technical^processing classification.
faceted^classification. colon classification

Title: Academic libraries
[academic, libraries]
academic ^ library
Library Science 6 Library academic^library

Title: Administration of technical libraries.
[administration,of,technical,libraries]
administration
technical^library
Library Science 6 library. special library 6.1 management