CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Leeds Beckett Repository

Leeds Metropolitan University Repository

http://repository.leedsmet.ac.uk/

Citation:

Guest, E. (2009) Library Cataloguing and Role and Reference Grammar for Natural Language processing Applications. **The 13th World Multi-Conference on Systemics, Cybernetics and Informatics: WMSCI 2009, July 10th - July 13th, 2009**.

This paper was submitted and accepted; however, it was not presented / published due to non-attendance at the conference.

Elizabeth Guest (2009)

The Leeds Metropolitan University Repository is a digital collection of the research output of the University, comprising citation records and, where copyright permits, full-text articles made available on an Open Access basis.

For more information, including our policy and submission procedure, visit the website at http://repository.leedsmet.ac.uk/

Alternatively, email the repository team at repository@leedsmet.ac.uk

Library Cataloguing and Role and Reference Grammar for Natural Language processing Applications

Elizabeth Guest
Innovation North, Leeds Metropolitan University
Leeds, LS6 3QS, UK

ABSTRACT

Several potential application of natural language processing have proven to be intractable. In this paper, we provide and overview of methods from library cataloguing and linguistics that have not yet been adopted by the natural language processing community and which could be used to help solve some of these problems.

Keywords: library cataloguing, PRECIS, Role and Reference Grammar, Semantics, natural language processing

1. INTRODUCTION

Solutions for problems such as automatic marking of free text answers, automatic summarisation, and automatic semantic indexing of web pages, would be likely to have a major impact on our lives. However despite decades of work such applications have proven to be intractable. It may be that significant steps forward could be make by looking outside the standard techniques that are currently being explored and incorporating ideas from other disciplines. To this end, this paper provides an overview of methods and techniques from other disciplines that have not yet been adopted by the natural language community and which could be used to help solve some of these intractable problems.

The three methods chosen are Role and Reference Grammar (RRG) [2], library cataloguing, and ULM [1], which is a semantic framework that attempts to both bridge the gap between syntax and semantics and handle ranges of meaning. Role and Reference Grammar is a linguistic theory similar to functional grammar. It is gaining a following in the linguistic community as various researchers apply it to a wide range of languages and plug the holes in the theory. This theory has attractions in that it has been shown to handle tricky grammatical structures in many different types of languages with little difficulty. It separates the most vital parts of the sentence from the modifiers (adverbs, adjectives, auxiliaries, and articles). This means that the core meaning can be extracted first and then the modifiers fitted in at a later stage. As long as the arguments and the verbs are in the correct order for English (subject verb object) then the sentence can be understood. It doesn't matter if (for example) Chinese students forget the articles, the sentence can still be parsed and the meaning extracted. An important advantage of RRG is that it is based around the notion of a CORE, which contains a predicate (normally a verb) and its arguments. The CORE is structured in such a way that it is easy to obtain the meaning from the sentence. More details of RRG are given in the paper "Parsing using the Role and Reference Grammar Paradigm", which has been submitted to this conference. Role and Reference Grammar includes a semantic paradigm based, like many others, on first order predicate logic. An investigation into semantic paradigms was performed and none were found that are really adequate to the task because they assume that words have distinct, well defined and non-overlapping meanings. In an attempt to solve this problem, a new semantic framework called ULM [1] was developed. This has not yet been thoroughly tested and will require refinement, but it is believed that it will be useful in tackling problems involving semantics.

The need for methods to catalogue items held in libraries has been an issue ever since libraries existed – more than 4000 years. Current methods enable someone searching for information in a library to find something about it relatively easily, particularly with on-line searching techniques. Librarians have already tackled many of the problems confronting those trying to get the semantic web up and running and it makes sense to see if they have anything to offer to make this task easier.

2. LIBRARY CATALOGUING

A library catalogue consists of entries containing metadata about each item in the library. The metadata includes all kinds of information about the item including the author(s), title, publisher, date of publication, edition, language, shelf mark, ISBN, price, physical description, type of material, key words, and much more. Metadata can be categorised into the following categories:

- Structural
- Indexing
- Administrative
- Descriptive

Descriptive metadata is the metadata used to describe the contents of an item and is the metadata most important for searching a catalogue. A library index can be built from the descriptive metadata and a subject heading list or a thesaurus and the latter ensure that consistency in the descriptive metadata is maintained. Note that a thesaurus in this sense is not a dictionary of synonyms but an augmented taxonomy. Today with the advent of online searching, an index is no longer required because the system simply retrieves all entries containing the entered search terms. However, this does not mean that thesauri and subject heading lists are no longer required. On the contrary they are very important in that they specify a controlled vocabulary that is used for the metadata entries to ensure that different words are not used to describe the same concept. This ensures consistency in the metadata, and improves search results.

Thesauri contain more complete information than a subject heading list as they contain both hierarchical and associative relationships as well as an alphabetical list of terms. Associative relationships contain all relationships that are not parent or child

relationships. They include sibling relationships. Some thesauri distinguish between generic parent-child relationships and part-whole relationships. Therefore a thesaurus can provide both an alphabetical list of terms and a map of the subject area. It also documents preferred terms and synonyms of these preferred terms are included in the alphabetical list. The choice of preferred terms is systematic.These preferred terms are useful for search as they provide for consistent retrieval of relevant items in the catalogue: key words in the catalogue entries will contain only preferred terms. By using the thesaurus, non-preferred search terms chosen by a user can be converted into the preferred terms. This means that no matter the terms found in the original item, retrieval by subject will be facilitated. The thesaurus is important for ensuring that items can be found even when someone searches using a different term.

3. AUTOMATIC LIBRARY CATALOGUING

There is a method of library cataloguing, PRECIS [3]. Which was discontinued during the 1980s becasue of the cost of manpower to do it. However, some librarians think it was better than the current methods, and would like to see it resurrected. It so happens that questions that someone using PRECIS has to answer are precisely those that Role and Reference Grammar is designed to answer. PRECIS is a method for generating an index of materials. It was developed in the 1960s and part of the process was automated even then. Its best known use was to provide the alphabetical subject index to the *British national* bibliography. The index entries consist of a few words that provide a summary of what the document is or what it is about. The process involves the indexer examining the document and making a brief summary of what it is about. This summary statement is analysed by answering the following questions

1) What happened? – the action
2) To whom or what did it happen? – the object of the action
3) Who or what did it? – the agent of the action
4) Where did it happen? – location
5) Are any of the concepts in the statement related in a whole-part relationship?

This produces a list of terms and the cataloguer then decides which should be the terms that appear in the catalogue. In general all terms will be lead terms. Only terms that are very general relative to the subject of the thesaurus or those that would generate myriads of entries are excluded.

An example (taken from the PRECIS book [3]) is

Subject: planning the planting of vegetables
String: (1) vegetables √
(2) planting √
(2) planning
Entries: **Vegetables**
Planting. Planning
Planting. Vegetables
Planning

Note that the ticks after the items denote lead terms. ‘planning’ is not a lead term because it is too general. The entries consist of several items which are in a specific order. The first line contains the lead term followed by ‘qualifiers’, which contains items higher up the list in reverse order. The second line contains the ‘display’, which contains items lower down the list in their input order. The numbers before each item in the list denotes its role. Roles are numbered as follows:

(0) the location of the action
(1) the object of the action
(2) the action
(3) the agent of the action

This denotes the filing order of the terms in the entry. They also play a role when the computer generates the entries from the string.

When the indexer has made and organised the list of terms, it is checked against a list held in a file. If the list already exists then the appropriate metadata is added to the catalogue entry. If it is not in the file, the list of terms is passed to the person in charge of the thesaurus. This person checks the string against thesaurus entries to make sure that only preferred terms are used. In addition, the thesaurus is updated if any of the terms are not yet in the thesaurus.

The thesaurus is an important part of the PRECIS process. It is used in the automatic generation of the index to generate ‘see’ entries which point the user to the preferred terms. It also generates ‘see also’ entries which point the user to narrower terms (child terms in the thesaurus hierarchy).

The hardest part of the PRECIS process is generating and organising the list of terms. This part includes the analysis of compound terms, which today is part of constructing a thesaurus. However, the process of generating the list of terms is remarkably similar to the ideas behind Role and Reference Grammar (RRG). Role and Reference Grammar is primarily concerned with the roles of the items in a sentence and contains methods for extracting these. In the roles are given the following names

(0) Periphery. Location is considered to be peripheral information.
(1) The undergoer (the object of the action)
(2) The predicate (the action)
(3) The actor (the agent of the action)

Everything within the design of RRG is to make the identification of the predicate, the undergoer and the actor within a sentence easy. Note that (1) can also be the subject of an instransitive verb, but this is no problem for RRG: it is just the ‘priviledged syntactic argument’. The PSA or priviledged syntactic argument is a device within RRG to identify the subject of an intransitive verb and to handle passive constructions appropriately. The only thing that might cause difficulty to RRG is the location, but this is really just a case of looking for locations in the PERIPHERY, which is not hard if the words are labelled correctly in the dictionary used for tagging. However, even this may not be necessary: a simple call to WordNet to see if the word comes under location may be all that is required. An examination of the preposition used to introduce the location will also provide clues.

When PRECIS was in use[1], much of the role of indexer was to apply world knowledge to the terms, especially when it came to dealing with compound terms. But if this is moved to the thesaurus, virtually all of the analysis of the statement of what the document is about can be automated

An advantage of building a PRECIS index is that an appropriate section of the index could be presented for the user to browse. This is almost certainly quicker that trying to interpret titles. In addition, information about the faculty in which the learning object was developed can be added, which will give the user additional information about whether or not the item may be useful.

4. Universal Lexical Metalanguge (ULM)

[1] For instance in providing the subject index to the *British national bibliography* 1971-1986

The Universal Lexical Metalanguage, ULM, is the outcome of a collaboration between a computer scientist and a linguist to find a better way of doing semantics. The result is a framework that is a combination of a lexicon and a knowledge representation system that has the following properties:

- strong link between syntax and semantics
- use of a universal semantic metalanguage
- Rich and expressive semantic structure (formal language) and explicit encoding of the argument of verbs to provide contextual information.
- use of fuzzy logic and fuzzy set theory to mimic how people use language
- explicit separation of objects, predicates, and operators

ULM provides a semantic knowledge representation based on meaning, and its theoretical stance allows us to develop powerful reasoning algorithms that enable it to be applied to many applications which require some analysis of meaning. For example algorithms for comparing sentences for similar meaning and reasoning with context can be applied to the problem of automatic marking of short free text answers. ULM has been presented to several people and some interest has been raised for its application to automatic translation (the UK security services), automatic marking, plagiarism detection (JISC), and for information retrieval within specific domains (several companies in Spain). These potential users see its potential to provide a more precise and elegant solution than existing algorithms which mainly base their search engine from a source word to a target word.
ULM is a schema consisting of several interlinking parts. This schema is based around two separate ontologies: one for predicates (generally verbs) and one for objects, which are arguments of predicates (generally nouns). Each predicate and each object is defined using a universal semantic metalanguage. This metalanguage consists of universal primitives and functions that will enable a full description of any language, and a mapping between languages.

The ontologies are linked to each other so that, for example it is possible to work out which prepositional phrases are arguments or argument adjuncts of the predicates, and which give other information such as time and place. Predicates and objects are linked together so that predicates are directly attached to their possible arguments and objects are attached to predicates that are applicable. These latter links can help to pin down the meaning of the predicate when applied to this particular object. This also builds in a framework for knowledge representation as it would be fairly easy to deduce from such a network what objects are used for. In addition we propose to link objects together into some kind of fuzzy conceptual ontology so that, for example, the word “meal” can be attached to “table”, “cutlery”, “food”, “drink”, “human” etc, so that a whole context can be derived (as each object is also attached to appropriate predicates).A schematic diagram of this schema showing how parts interact is given in Figure 3.

In relation to the definitional apparatus, each prime will be at the top of its own hierarchy defined by a set of hyponyms derived from these semantic primes using universal functions applied to the intervals and a formal definitional language. These hierarchies each define a separate domain and provide a disconnected but well defined set of domains, based on universal primitives. Each derived word will inherit some or all of the intervals from the relevant prime and these intervals will describe the range of meaning of this word. These derived words are language specific because the primes are language specific. However, the metalanguage with which they are described is universal.

Words that contain ideas from more that one domain will sit between these domains Each word derived in this way will also be language specific. However, it will be possible to map concepts between languages by looking up common sets of intervals used and common definitions of words.

A schematic diagram of the predicate ontology is given in figure 1 and a more detailed view of how this would work in practise is given for English verbs in Figure 2

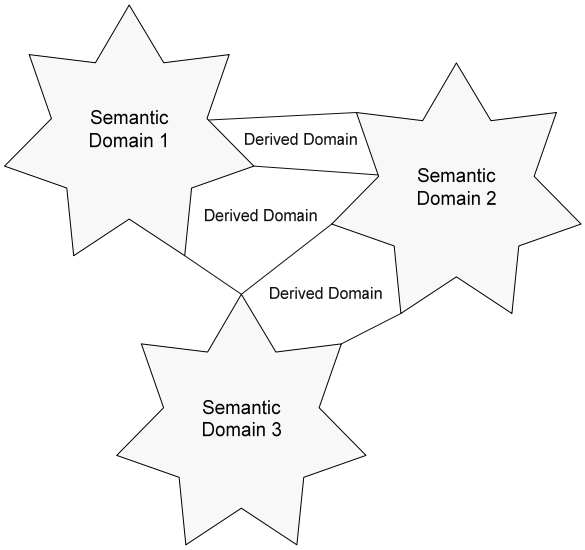

Figure 1: A schematic diagram of the predicate semantic domain.

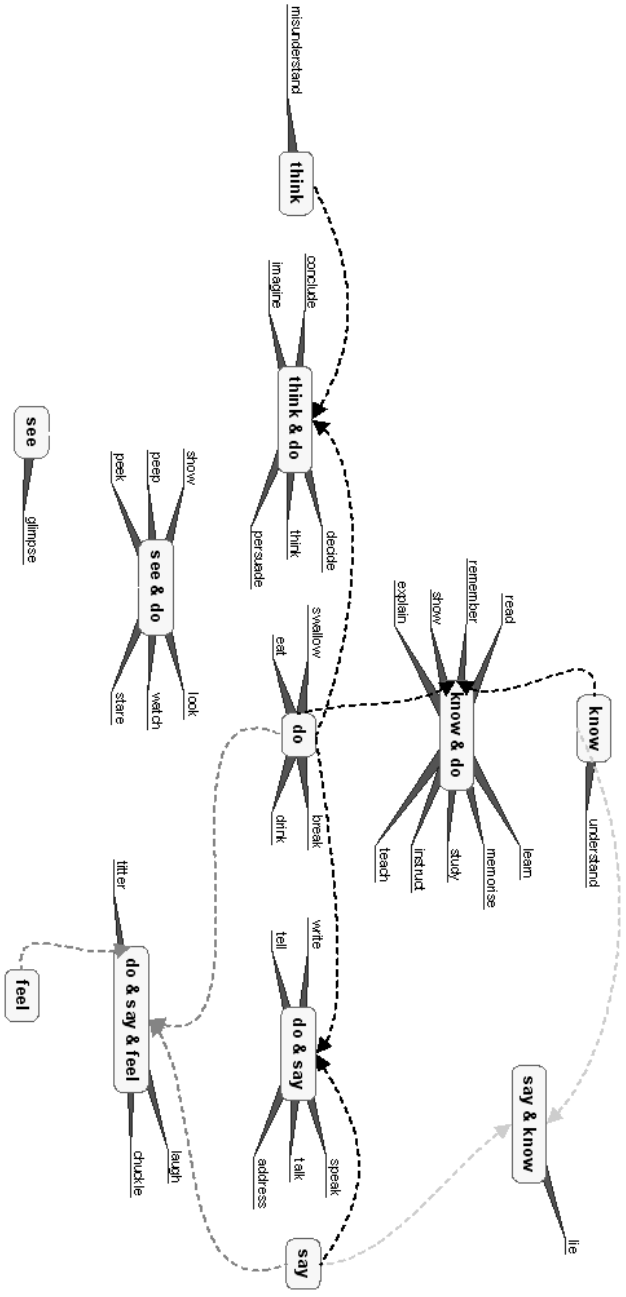

Figure 2: A partial ontology of English verbs. The arrows show that words that are derived from more than one prime inherit the intervals from all relevant primes.

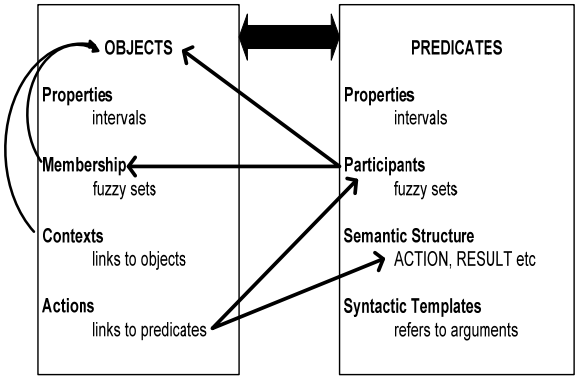

Figure 3: Overview of the new semantic framework

5. Conclusion

The methods briefly described above could be applied to several application including automatic library cataloguing, the semantic web, and automatic marking of free text answers. Of these the application of automatic library cataloguing has been described in most detail, but this clearly has implications for the semantic web. If automatic library cataloguing is possible, we are not far from being able to catalogue web pages in a meaningful way. Methods for automatic marking of short, free text answers would use the methods of automatic library cataloguing augmented with ULM to handle semantic information. Automatic cataloguing would also benefit from the use of ULM, but given a suitable thesaurus (such as the Library of Congress Subject Heading list), it should be possible to manage with existing semantic repositories such as WordNet.

6. References

[1]. Guest, E. and R. Mairal Usón, "Lexical Representation Based on a Universal Metalanguage". **RAEL, Revista Española de Lingüística Aplicada**, 4 2005 pp. 125-173.

[2]. Van Valin, R.D.J., **Exploring the Syntax-Semantics Interface.** Cambridge University Press.2005

[3]. Austin, D., **PRECIS: A Manual of Concept Analysis and Subject Indexing.** The British Library 1984