The Use of Lexicons in Information Retrieval in Legal Databases

JC Smith
FLAIR Project
Faculty of Law, the University of British Columbia
1822 East Mall, Annex 1
Vancouver, BC V6T 1Z1
jcsmith@flair.law.ubc.ca WWW:http://flair.law.ubc.ca

Abstract

This paper reports on the mid-beta stage of development of the Flexicon system and describes how a lexically designed search engine in a domain specific database such as a large database of legal cases, can provide better relevance ranking than best match search engines, and at the same time permit an information need to be formulated using multiple word concepts, phrases and items, which is the one great advantage that an exact match Boolean search engine has to offer over best match engines. In addition the paper sets out in detail the lexical structure of the Flexicon system. The first tests on a legal problem, comparing the Flexicon system with two best match systems, indicate that a lexically designed search engine and database has the potential for a substantially higher level of precision than best match search engines.

1 Introduction

It was just fifty years ago that John Bardeen, William Shockley, and Walter Brattain invented the transistor. This year, 1997, is the year that HAL, the computer in the Arthur C. Clarke and Stanley Kubrick film, *2001: A Space Odyssey*, that could think, talk, see, feel, read lips, and go "berserk" was supposed to have come into consciousness and be operational. [Garfinkel] The chip has undergone transformations through miniaturization and microelectronics beyond Clarke and Kubrick's wildest imagination. Yet. a computer such as they conceived of in their science fiction future, at this point in time and at least for some time to come, looks next to impossible. Some, however, still have hope. [Stork] No doubt the "true believers" in *real* AI will have had their optimism renewed when, a few weeks ago artificial intelligence research passed a new and important milestone with the victory of IBM's Deep Blue over Garry Kasparov, probably the greatest chess champion in history. Yet, on the other hand, the most powerful computer still cannot manage natural language at the level of a two year old child. Possibly there are some deficiencies in the paradigms we use in AI. If our AI paradigm is misconceived it can result in misdirection of effort and a waste of scarce resources. An alternative, albeit more modest, paradigm could well point research and development in different and more fruitful directions.

ICAIL-97, Melbourne, Australia © 1997 ACM 0-89791-924-6/97/06.$3.50

Terry Winograd wrote that, "In the tradition of artificial intelligence, we project an image of our language activity onto the symbolic manipulations of the machine, then project that back onto the full human mind. But these projections....systematically eliminate dimensions, thereby both simplifying and distorting.... [W]e all too easily dismiss the concerns of human meaning that make up the humanities, and indeed of any socially grounded understanding of human language and action....we lose sight of the tacit embodied understanding that undergirds our intelligence." [188] It was ten years ago this year that the First International Conference on Artificial Intelligence and Law was held in Boston. During the past five conferences, and the now many issues of the Journal, we have reports of many different kinds of research, from a variety of different perspectives, and we have the opportunity to examine our own progress in the light of Winograd's concern

The University of British Columbia Faculty of Law Artificial Intelligence Research (FLAIR) Project had its origins in a three year cooperative project between UBC and IBM (Canada). FLAIR, from the beginning, pursued a theoretical vision, having as its objective the integration of a particular paradigm of artificial intelligence with psycho-linguistics, and legal theory. The key underlying factor of this perspective was the relationship between the discourse of legal doctrine (as signifiers and chains of signification) and the discourse of the material facts of life (as the signified) to which it is applied. We, at FLAIR, believe that the key to understanding how legal reasoning functions lies in understanding that relationship. Our particular psycho-linguistic perspective is based on the theory of language of Ferdinand de Saussure and the further developments of those insights about language (often referred to as postmodernism or poststructuralism), in the works of such social theorists as Jacques Lacan [1977. 1988, 1993] and Jacques Derrida [1976, 1978]. Our first project was the development of a methodology for simulating case-based reasoning in the computer through representing the ordering and relationships of the underlying teleological or goal structures of legal reasoning, and formulating the dynamics of decision-making in the form of deep-structure rules. [Smith & Deedman, 1987; Deedman & Smith, 1991; Smith, 1993; Deedman 1994; Kowalski, 1991; MacCrimmon, 1989] We characterize this perspective as teleo-analytic jurisprudence [Smith, 1976; Coval and Smith, 1982, 1986], since it entails an analysis of the teleological structure of the law and the processes of legal decision making.

Early in our research and development of our expert system/case-based reasoning deep-structure technology, we

recognized that this methodology, while fruitful in confirming some of our theories about legal reasoning and discourse, would have little practical application because of the amount of time and resources involved in constructing and maintaining such systems. At the same time we began to recognize that our particular theoretical framework might offer a possible solution to some of the problems of the management, representation, and retrieval of legal information from electronic databases, and so we decided to take our research in this direction.

On April 29th of this year, the Oklahoma Supreme court mandated by court order vendor-neutral public domain citations for all of its decisions for its entire body of case law reaching back to 1890, and has provided a "converter" tool to convert West citations to the public domain citations, so that users may "cut and paste" the citation into their documents. The rule requires the continued use of the National Reporter System citation as a parallel cite. The Oklahoma courts have, as well, made the commitment to make freely available all Oklahoma decisions since 1890 on its Web site in a searchable format. By the end of 1997 it is likely that at least 10 states will have made available public domain vendor-neutral citations for their cases. The Canadian courts have made the decision to implement a vendor neutral system, and practically all cases now in Canada, appear with paragraph numbering. As more and more cases, statutes, and regulations become available in electronic form, and as more and more courts adopt paragraph numbering and vendor neutral citation systems, and accept citations to electronic databases, the need for a user friendly highly efficient search engine becomes more pressing.

There are only two basic kinds of search engines, Exact Match—Boolean, and Best Match—Non Boolean. A Boolean search engine permits the use of multiple word groupings in the search query, but does not return the documents ranked as to how well they match the information need which the query represents. Non-Boolean—Best Match search engines return the documents with some semblance of relevance ranking but they do not permit the use of multiple word groupings and phrases in the search query. The user is limited to using single words. Each of the two kinds of search engines are structured very differently, and consequently, since the existing search engines cannot be combined, one must choose between relevance ranking or multiple word queries, even though both are essential for efficient document search and retrieval. Certain "fixes" can be made to each kind of search engine to ameliorate the lack. Limited statistical-based relevance ranking can be added to exact match search engines, and a limited multiple item list can be added to best match systems to allow specific frequently appearing phrases to be recognized as single items. [Croft, Turtle, & Lewis, 1991]. These considerations led us to create Flexicon.

The Flexicon system is a lexically designed and structured search engine for domain-specific vertical information markets such as law and medicine which combines both relevance ranking and multiple-term queries, permitting a precision of recall far exceeding any existing search engine on the market. The lexical structure of the Flexicon system permits the user to approach the information through hierarchical doctrinal lexicons which function like the table of contents of a book, and alphabetical lexicons which function like indexes. The lexicons represent legal knowledge in the form of a quadrant of legal concepts, factual terms and phrases, case citations and statute references. Legal information in print generally contains lexicons of these four items, often in both alphabetical and hierarchical order.

The development of the Flexicon system has gone through a number of stages. The first stage was proof of concept consisting of a functioning system on a database of about three hundred cases. [Smith & Gelbert, 1991, 1992]. The second stage was the development of a prototype on a database of about twelve hundred cases, which was then tested against a Boolean engine on the same database. [Smith & Gelbart, 1993] The alpha implementation consisted of a database of about thirty three thousand cases, consisting of the California Third and Fourth Series of reported decision, which requires two CD-ROM disks to contain the system. [Smith , Gelbart, et al, 1995]. In alpha testing we explored the power of the system to recognize multiple word concepts and phrases. At present, we are developing a beta version of the system. We have located the types of multiple word items which the alpha version fails to recognized, and are presently in the process of implementing the solutions. The new beta version will have an entirely new user interface design.

2 Lexicons

In recent years, as people working in the fields of cognitive science, linguistics, psycholinguisitics, computational linguistics, artificial intelligence, natural language processing, and information technology have become increasingly aware of the complexity of natural language, lexicons and lexicography have become topics of widening interest. [Atkins & Zampolli, 1994; Wilks, Slator & Guthrie, 1996; Guo, 1995; Boguraev & Briscoe, 1989] In 1989, Donald Walker and Antonio Zampolli wrote that, "the development of large lexical knowledge bases has emerged as probably the most urgent, expensive and time-consuming task facing linguistics, computational linguistics, and artificial intelligence." They go on to point out that the number of specialized workshops and conferences in the field justifies the conclusion that computational lexicography and lexicology is emerging as "a discipline in its own right." [xiii] We consider any list of words which has been gathered and arranged according to any kind of structure which serves a particular purpose, to be a lexicon. Thesauruses and dictionaries are particular kinds of lexicons. A dictionary is a lexicon which defines particular words in terms of other words in the same language, or in words from a different language. Indexes of books are lexicons which represent the content of a book in terms of the significant words, phrases, and concepts arranged in alphabetical order, which facilitates the process of quickly finding relevant information in the volume. The table of contents of a book is also a lexicon in that it is a hierarchically ordered list of words and phrases which represents the content of the book.

The lexicon is the point where psycholinguistics and computational linguistics and standard familiar book-based research methodology converge in that the lexical component of language is common to the human, to books, and to the computer. [Beckwith, Fellbaum, Gross & Miller, 1991] Most of the lexical properties of language can be represented in computer-based lexicons. As humans we are used to using lexicons or dictionaries which are ordered alphabetically. Alphabetical organization is a widely used

computer lexicon function. We are also familiar with ordering information hierarchically. Hierarchies are widely used in computational linguistic technology to organize large databases of information. Computer scientists use hierarchies to create inheritance systems whereby the subordinate can be assumed to have inherited or to possess the properties of the superordinate. Synonyms can be linked in computer-based lexicons. The use of lexicons permit us to create lexical matrixes in the computer.

3 An Alternative Paradigm for AI

One of the more striking manifestations of intelligence is the capacity of humans to carry out a difficult task by creatively transforming the problem into the form of a number of simpler tasks, and then reconstructing it in a more complex form in order to achieve the desired objective. The mathematician Zdzislaw Melzak calls this process the *bypass principle* which "is a way of dealing with complexity or with difficulty by means of a bypass which promotes a transport or a passage or the solution of a problem in a three-stage process whose first and last stages are each other's inverses".[Melzak]

When language initially evolved, it existed first as an oral tradition. Prior to the invention of a written language, passing down the history and culture of a group was a difficult task. In each generation, people had to acquire the information from the elders, memorize it, and pass it on to the next generation. With the evolution of writing, the concepts and corresponding sounds of the language were encoded into a visual symbolism by inscription in material ranging from stone to papyrus, and then decoded in the process of reading. The literate tradition took a revolutionary form with the development of the printing press. With the development of digital technology, we enter a new era in the symbolic representation of human thought.

Words can be represented in the form of clicks of a telegraph key, which in turn became electrical impulses, which, after passing through a wire are turned back into clicks, which are interpreted as words. Many of the more complex actions which best reflect human intelligence follow a pattern whereby a very difficult task is broken down into or transformed into a representation in a much simpler form. In this form, a set of easy task can then be carried out. When, through a process of inverse transformation, the representation is returned to its original form, the original tasks is completed. This gives us the formula:

DIFFICULT TASK to SIMPLIFIED TRANSFORMATION to a set of EASIER TASKS to an inverse SIMPLIFIED TRANSFORMATION to the COMPLETED TASK.

The case-based reasoning deep structure methodology which was developed at FLAIR was based on the above formulation. [Kowalski] An area of law is chosen which presents the decision maker with a hard case, and thus a *difficult task.* The facts and complex legal doctrine of each of the legal cases which constitute the case-base of the system are reduced by a *simplified transformation* to an ordered set of goals. The facts of a legal dispute falling within the domain of the system are ascertained by a set of questions designed to identify the goals which the controversy brings into conflict. The case-based reasoner then seeks by pattern matching to find the predominant ordering of these same conflicting goals in the database, *the set of easier tasks.* That ordering returns the corresponding set of cases which have decided the ordering of the goals in dispute, the *inverse simplified transformation.* An algorithm assigns a weight to each case depending upon such factors as the age of the case, the jurisdiction, and the level of the court. The system then calculates an outcome, (the plaintiff or the defendant wins) in terms of a percentage, *the completed task*, and returns the relevant cases. Thus the relationship between legal doctrine (*the chains of signification*) and the facts of material life (*the signified*) are represented in the form of deep-structured patterns of consistency in the way the goals of the law have been ordered in prior decisions, (*the associative relationships which relate the signifiers to the signified*).

Lexicons are representations of discourses. A dictionary is a representation of a language, and a table of contents and a set of indexes are representations of a book. Lexicons have an "architecture" which incorporates meaning from the discourse or texts which they epitomize. The architecture of legal lexicons embodies to some degree the teleological relationships between the material facts of a legal dispute and the doctrinal structure which is used to justify the decision. The Flexicon system in development at FLAIR has sought to use lexicons as *creative transformations* to solve some of the *difficult tasks* related to the retrieval of information through the process of, permitting the computer to carry out a complicated series of *simple tasks*, which can then be *inversely* returned in a form wherein the *difficult task* has been completed. We set out to develop a new approach to document retrieval in large domain specific databases, such as those we find in law. The method we chose was to represent the databases, and the documents which constituted them, in terms of sets of lexicons organized alphabetically, hierarchically, and in terms of frequency. These various lexicons and sets of lexicons are all interrelated.

The databases in a Flexicon information system exist electronically in three forms. The first is the lineal text of the original document, the second is the entire database in the form of sets of multiple lexicons, and the third is each document in the database individually in the form of a set of lexicons. These three forms of the data are interrelated into a lexically-based matrix. Each document is retrieved in the form of a FlexNote where the various kinds of lexicons which constitute the content of the document are presented to the user in terms of frequency of terms normalized over the content of the entire database. The relevancy of a case can often be determined at a single glance. The FlexNote, therefore, constitutes a machine created abstract of the document. One can then hypertext from each item in the document profile to each of its occurrences in the document, or go directly to the full text. The objective of the Flexicon technology is to have the machine, as far as possible, process the raw data, create the lexicons, automatically classify the subject matter, insert hypertext links, and recognize header or title formation.

4 Making the Invisible Visible

Information in the form of books has significant advantages over information in electronic form. The contents of a book is observable on its face through table of contents, indexes, summaries, and one can skim the contents by quickly paging through and glancing at each page. Imagine, however, how difficult it would be to obtain

information from a book that had neither a table of contents nor a decent index. It would be like looking for a needle in a haystack. Looking for a needle in a haystack is not an inappropriate metaphor for information retrieval in large databases. Information retrieval in an electronic database is a much more difficult task than the retrieval of information from books in that the information in an electronic database is not visible until it is actually retrieved and transformed into the visual symbols of the alphabet. The difficulty lies in that you cannot see the information until you locate it, and you cannot locate it unless you can see it.

The process of information retrieval, whether from a book or an electronic database, follows the same pattern. The information needs of the researcher must be simplified in the form of a transformation which will represent the information need. This creative transformation, whether a member of a conceptual hierarchy or a list of words, is then matched with the corresponding representation of the content of the book or of the electronic database. When the relevant match is found, the simplified representation is transformed back into the relevant text. Tables of contents and indexes are simplified representations of the contents of a book.

Information retrieval becomes more efficient in databases rather than books because of the huge amount of information which can be inputted and stored, and the speed with which that information can be accessed, manipulated, and retrieved. Simple tables of content and indexes such as those found in books have diminishing returns as the size of databases increase, and are time consuming and expensive if manually created. The difficult task in doing research with electronic databases is to design an information system which will automatically generate hierarchical conceptual structures, alphabetical indexes for words and phrases, and sets of lists of significant factors out of the content of each document, in a domain specific database. These structures, indexes, and lists would permit the user to examine the content of the database and locate the words and concepts which match the information needs. The Flexicon technology consists of a set of libraries and tools which can be used to create these lexicons in specific domains such as law or medicine, or any other areas which lie within specific subject fields. In this manner the invisible is made visible.

5 Theoretical Foundations.

The Flexicon sytem was constructed on the basis of the linguistic theory of the Swiss linguist, Ferdinand de Saussure. Saussure demonstrated that metaphor and metonymy are the driving force of language. [161-173] One of the essential problems of language is the nature of the relationship between sound images and concepts, and between words and concepts (signs), and what they refer to. [66] Saussure drew a fundamental distinction in the way language functions, between what he called the *signifiers* and the *signified.* Signifiers are related to each other in terms of chains of meaning or signification. These chains take the form of regular or orderly collections of statements, propositions, or doctrines, which are interrelated in terms of interdependencies so far as meaning is concerned. The relationship between signifiers is *syntagmatic.* "Between the syntagmatic groupings....there is a bond of interdependence; they mutually condition each other." [128] Syntagmatic groupings of language at one level of discourse are associated with other levels of discourse as the signified. The relationship between signifiers and the discourses which function as the signified is neither logically necessary nor semantically fixed, but rather is more metaphorical. Saussure referred to the relationship between the chains of signification and the signified as an *associative* relationship.

The nature of the relationship between the discourse of legal doctrine (the Law) and the discourse of the observations, information, and data of everyday life (the Facts) is one of the most difficult issues facing the practitioners of AI and law, working in the area of case-based reasoning and legal expert systems. If the relationship was fixed and determinate, our task would not be nearly as difficult as it is. Since the relationship is indeterminate and shifting, simulating legal reasoning in the computer presents an overwhelming challenge. Viewing the discourse of legal doctrine as signifiers and chains of signification, and the discourse of the material facts of life as the signified furnishes us with a useful conceptual tool for examining the relationship between what we call the law and the facts.

At times the lawyer starts with a set of facts as a given, and attempts to find sets of legal doctrine which will, when applied to those facts, produce the result desired by the client. Take, for example, a set of facts in which a hospital wrongfully posts information that a particular nurse has had a highly infectious disease, and people assumed that she had AIDS rather than infectious hepatitis. The cause of action could be drafted in any one or more of the following legal actions: negligence, breach of contract of employment, defamation, or invasion of privacy The law of contracts, the law of defamation, the law of privacy, and the law of negligence —each are doctrinal sets which are made up of chains of signification. This discourse is normally fairly abstract at the purely doctrinal level, and doctrinal arguments are generally fairly concise because they are frequently circular within their own frame of reference. The discourse which would be used to describe the material facts of the situation involving the nurse and the damage she suffered, constitutes the signified. The essence of a legal argument is to persuade the judge to select as the chains of signification, a particular doctrinal analysis which when associated with the facts of the case, a conclusion will follow as a matter of course, which will be favorable to her client.

On other occasions, the lawyer has located the appropriate legal doctrine which might produce the desired result, and seeks alternative fact situations where that same doctrine has been applied in order to use as precedents. For example let us take the case where a young female university student is returning to her car in the university parking lot, and while walking down an unlit path bordered by shrubs and bushes is sexually assaulted. The legal issue is as to whether or not an occupier of land owes a duty of care to invitees or licensees to take precautions to protect them from the intentional wrongful act of third parties. There are a wide range of possibilities which could give rise to this same legal issue, and consequently would be relevant cases. The nature of the premises, the occupier's business, the location of the wrongful act, the nature of the wrongful act are all variables which can change and still remain within the range of relevancy. The separation of the discourse of legal doctrine from the discourse of the material facts of life, therefore, not only has a theoretical

justification, but a practical application regarding how we structure and represent legal knowledge in the computer.

The Flexicon system is based on the assumption that the discourse which constitutes many discrete subject areas will be made up of a number of different syntagmas, chains of signification, or doctrinal discourses, which are associated with other sets of discourses of a more factual nature. Thus the discourse of medicine, for example, contains many sub-discourses such as anatomy, pathologies such as diseases and injuries, symptoms, and treatments. Where large databases consist of documents which fall within a particular subject area, where the discourse of the area is made up of sub-sets of separable kinds of discourses, the Flexicon system is designed to separate these sub-sets, and present, represent and retrieve them in the form of multiple lexicons — hierarchical to show subordination and superordination (hyponymy), and alphabetically for ease of location of specific terms.

6 The Lexical Structure of the Flexicon System

The Flexicon system consists of two basic kinds of lexicons, static and dynamic. The static lexicons are fixed, while the dynamic lexicons are created during the information retrieval process. The major function of the dynamic lexicons are to reveal the content of the database in a lexical form centering around a particular word or concept. The user enters the word or concept, and the system then creates a lexicon of terms and phrases which include that word or one of its forms, either at the beginning, at the end, or inside the phrase, and returns the lexicon alphabetically ordered. Words which are entered go through manipulations and transformations such as stemming, and the intelligent elimination of punctuation and spaces in order to produce a standard form. The system, as well, contains master lexicons which are used in the creation of the lexicons which will constitute the particular database. The master lexicons are more or less complete but their counterparts in the particular databases contain only those items in the master lexicon which are to be found in the particular databases.

The interface between legal doctrine (the signifiers) and the material facts (the signified) is the focal point for representing legal reasoning. Each decided case is an application of a set of legal doctrines to a factual situation, and each case citation is a representation of signifiers applied to a signified. Phrases which combine both doctrinal and factual discourse often embody the relationship between a signifier and the signified to which it is applied. The words which constitute the phrase, **the duty owed by an occupier to an invitee**, are primarily signifiers. In the context of legal discourse, the words **University, student, and adequate safety measures** are primarily a part of the discourse of the signified. The phrase, **the duty owed by the University to a student to provide adequate safety measures** contains signifiers applied to the signified, and it "encrypts" teleological signification. The capacity of the dynamic lexicons to retrieve multiple word phrases which contain both the doctrinal signifiers and their signified is one of the most powerful features of the Flexicon system.

As well as being able to form a search query from the lexicons, the user can form a profile of the entire information need and retrieve documents in order of similarity of the document lexical profile with the information need lexical profile. The relevancy order of the returned documents can be significantly improved over other search engines through the weighting given to each lexicon and to individual items. There is no particular limit on the number of lexicons which can be used in the Flexicon retrieval system.

In the ongoing development of the Flexicon information system we are in the process of creating a set of tools for automating, as far as possible, the creation of lexicons from information available in electronic form. These tools will include concept development tools for the creation of hierarchically structured conceptual lexicons from domain specific texts, elimination lexicons, concordance programs, frequency measurers, non-specific or general purpose phrase lexicons, and special forms of word lists.

The Flexicon system takes an entire database of cases in electronic form and separates the text into five groups. It first eliminates the *noise words* such as *the, and, or, in, at* etc., but it does not eliminate them altogether. Through parsing functions, it retains them where they function as a part of a phrase. The system attempts as much as possible to recognize and retain whole phrases. The remaining text is divided into legal concepts (including single terms and phrases), factual terms and phrases, the names of cases, and statutory references. The concepts are organized both hierarchically as well as alphabetically, and statute citations and references are organized in the familiar way, alphabetically by statute within each jurisdiction, with the section references in the numerical ordering of the sections and sub-sections. The factual and case lexicons are in alphabetical order. The quadrant of concepts, cases, facts, and statutes is the organizational structure of the legal knowledge as represented in the machine.

7. Concept Lexicons

The most important Flexicon system lexicon is the hierarchical concept lexicon. It is hierarchically structured in terms of subject areas of the law, with each subject breaking down into several levels of sub-categories. The creation of this lexicon has been expensive, and time consuming, and is still yet not complete. The concept lexicon recognizes synonyms and alternative word forms. The distinction between what is conceptual and what is factual is often difficult to make. We have tried, as far as possible, to follow and use the linguistic theory of Saussure. Words which have little meaning standing alone, but require a conceptual doctrinal context would generally be classified as a legal concept, where words which have a fairly clear meaning outside of a legal context are generally considered not to belong in the concept lexicon even though they appear frequently in legal discourse. The fact that a particular term appears frequently in the context of a particular area of law, is not sufficient to warrant its classification as a legal concept. The term *gratuities*, for example, may appear frequently in the discourse of a particular area of tax law, but would not be included in the tax part of the concept lexicon.

Some words have both a doctrinal meaning in a particular legal context as well as an ordinary language meaning, and thus pose some difficulty. One way we have been able to deal with this kind of ambiguity is to use, where

possible, a multiple word phrase including the term, which eliminates the non-technical meaning. Thus we might get better results using the multiple word item *tax fraud*, rather than *fraud*, in the tax part of the concept lexicon, if the frequency of the multiple word item justifies its inclusion. There are certain legal concepts having a dual meaning where the legal concept must be kept as a single word. An example is the concept of **consideration** in the law of contract which appears in a variety of phrases having nothing to do with the doctrine of consideration, such as **taking into consideration, under consideration, judicial consideration**. We hope that this kind of problem eventually will be further alleviated through automatic subject classification of documents.

The second factor used in deciding inclusion or exclusion from the hierarchical concept lexicon is whether or not it appears sufficiently in legal discourse to warrant inclusion as an item for formulating a profile of the information need. There may be some obscure and antiquated terms which can be found in legal dictionaries but might not be included in the hierarchical concept lexicon. These are not lost, however, as the system will automatically look for a term in the fact quadrant if it cannot find it in the concept lexicon.

The third factor which we find useful in deciding on inclusion or exclusion is the ease with which an item can be fitted into the hierarchical structure. If one has a problem in locating an appropriate place in the doctrinal hierarchy then this is a factor in concluding that it does not belong in the concept lexicon.

Our concept lexicon is being developed by human expertise as automatic or machine concept recognition, thus far, we believe, lies outside of what is technically possible. We have developed a concept tool which facilitates the preparation of the hierarchical concept lexicon. Our concept hierarchical lexicon searches on synonyms as well as the root concept. When a concept is entered into the concept hierarchy through the concept tool, the synonyms are listed at the same time. Having the system automatically search on the synonyms, however, takes great care in that many synonyms have different alternative meanings. Thus **board** and **plank** are synonyms in the context of lumber, but they are not synonyms in the context of corporate organization, and a **plank** in the party's political platform has nothing to do with lumber.

All concepts, when entered, are given one or more of a set of three properties, *stemmable*, *movable*, and *separable*. If a concept is *stemmable* it means that the suffixes can be varied. For example, **rape** if marked as *stemmable* will find **rape, raped, raping**, and **rapes**. *Movable* means that the ordering of the words in a concept phrase can be rearranged. **Negligent solicitor** will be found in the phrase, *the judge found the* **solicitor negligent**, if the concept is *moveable*. *Separable* means that words can come in between concept phrases. **Negligent solicitor** will be found in the phrase *The defendant was found to be* **negligent** *while acting as* **solicitor** *of the plaintiff* if the concept is *separable*. The power comes into play when these attributes are combined. If **negligent solicitor** is marked *stemmable*, *movable*, and *separable*, it will also be found in *The* **solicitor** *was found liable in* **negligence**. Noise words such as **of** and **the** which occur in concepts are ignored only when the concept is marked as *separable*. Thus if the concept **impaneling of jury** is marked as *separable*, it will also recognize **impaneling jury, impaneling the jury,** and **impaneling of the jury.** When the concept phrase is marked *separable*, any word can come in between the concept words, not only just noise words. For example it will also find **impaneling the grand jury.** Everything in the concept lexicon is case independent. Concepts will be found whether they are upper case, lower case or a combination of both. A hyphen is treated as a hyphen or a space.

The many kind of ambiguities and inconsistencies in both the technical doctrinal discourse of law and the ordinary language of factual discourse normally do not negatively effect relevance ranking in the Flexicon system as any one or two individual items seldom carry enough weight within a well formed Flexicon information need profile to significantly effect search and retrieval.

8 Fact Lexicons

The Fact Lexicon is very different from the hierarchical concept lexicon. The concept lexicon is carefully constructed to recognize through a sophisticated pattern matching, each legal concept in the various forms which it might be found. The Fact Lexicon, on the other hand, is fundamentally a default lexicon. The Fact Lexicon is made up of fact words and fact phrases. Fact words are every word in the database other than concepts, statute citations, cited cases, and noise words. Fact phrases are fact words that appear next to each other with or without noise words in between. Within the Fact Lexicon terms are classified as modifiers, joiners, and fact words. Words that one would never want to use as a search term even with other words around it are eliminated as noise words and consequently do not appear in the Fact Lexicon. Modifiers are words that would be seen only in relationship with other fact words. You would never use it as a search term standing on its own. Modifiers are words that have little specific meaning on their own in isolation but need other words in order to give it a significant meaning.

Fact phrases are constructed using a simple grammar such as:

<fp> = fact word
<fp> = <fp> fact word
<fp> = modifier <fp>
<fp> = <fp> joiner <fp>
<fp> = modifier joiner <fp>

and so forth. At our present state of development Flexicon does not recognize every single phrase which would have significance as a profile item. We have located the kinds of phrases we are still missing and have designed solutions which we are now in the process of implementation and testing.

The Fact Lexicon is one of the most powerful features of the Flexicon system in that it permits a user to enter a term and to create a dynamic lexicon of all of the phrases in the database which contain that term in some form or other. This permits the user to locate phrases which the database contains which the user would probably have never thought of without this aid. Also, however, it presents some of the most challenging problems. For example **and** is an important *joiner* word for a phrase such as **Adam and Eve.** At the same time, however, you do not want it to pick up **going to the park and playing with the ball.**

9 Using Lexicons to Form a Profile of an Information Need

The earlier example of a typical legal problem was as follows:

The plaintiff, a university student enrolled in an evening class, had to walk down an unlit path bordered by shrubs and bushes, in order to reach the parking lot where her car was parked. When proceeding to her car, after her class, an assailant dragged her into the bushes and sexually assaulted her.

This statement of the problem is also a version of an information need. A good Boolean representation of it would be:

university college/p student and ((sexual! /2 assault) or rape! and (foliage or shrub! or bush! or tree) and security and negligen!

A best match natural language query would look something like the following:

Does an occupier owe a "duty of care" to an invitee or licensee to provide security or other safeguards on the occupier's premises against assault by a "third party" or otherwise provide a safe environment?

A Flexicon information need profile is constructed by selecting items from the lexicons. In most cases a user would generally start with the law. The concept hierarchical lexicon will permit the user to select the appropriate areas of the law and highlight in the corresponding dialogue box the items which the user wishes to use to form the profile of the information need. The system will also provide the user with a dynamically created lexicon of all forms of that term, or phrases in which it is to be found, and present them to the user in alphabetical order. Appropriate items would then be highlighted in the list for inclusion in the information need profile. Flexicon automatically behind the scenes includes all of the synonyms of that concept. The default weighting of medium can then be altered for any term where it is appropriate.

The next step would be consider the factual script to which the legal doctrine will be applied. The user can build up a profile of the factual elements of the script by entering in a dialogue box, a core factual term. The system will immediately create a dynamic lexicon of all of the forms and phrases which appear in the database containing a form of that word, or a phrase in which a form of it occurs. The plaintiff, in the above set of facts, will be a student, so one can enter the word **student**, and click on **Lookup**. The returned dynamic lexicon contains 604 alphabetically ordered references from which one might select, **attacks by non-students, college students, female student, protect students. Lighting** will produce a lexicon of 72 terms, **security**, 676 terms, and **campus** 136 terms. While these lexicons seem large, a user is able to page down through them very quickly.

At this point the user has a substantial profile of the facts of the particular case. The capacity of the system to create dynamic lexicons around specific terms permits one to enrich the profile to cover cases which raise the same legal issues but in a different factual context.

The lawyer generally creates a script out of the set of facts or story which gives rise to the legal issues in order to locate cases which may be somewhat factually different, but raise relevantly similar legal issues. The profile can be broadened to cover similar situations, such as other kinds of educational institutions, other kinds of wrongful acts, and other kinds of failures to remove different kinds of risks. Thus a lookup on the term **parking** will produce a dynamic lexicon of 489 terms in the California database from which one can select items such as **restaurant parking lot, parking garage, shopping mall parking area**, etc. all places where a criminal assault or a robbery might take place. Thus the user can form a profile of not only the particular set of facts, but a range of factual situations giving rise to the same legal issues.

The user now would be ready to do the initial search. The Flexicon system would then retrieve a large number of cases, ranked in terms of relevancy in regard to the problem profile. If one went through these cases in order of relevance, by looking at their FlexNotes, viewing the text and paging down to the facts, one would find that almost all are relevant to our legal problem. In the case quadrant of each FlexNote, the user will see the names of a number of cases, appearing more or less frequently. If one hypertexts down on each occurrence of these cases, one would soon discover that they are considered by the judges, as leading authorities for these kinds of problems. As the user locates these cases, highlights them in the FlexNote of a case the user can add them to the problem profile. About eleven or so cases will be found more frequently cited in the list of returned relevant cases, and so now ought to be added to the problem profile in the case quadrant, as leading and frequently cited cases. A similar examination of the statute references would be carried out next to locate any statutory provision which is cited in a number of the relevant cases. If one knows the name of one party to a case and wishes to look at the case or add it to the problem or information need profile, the user would enter the name in the dialogue box and do a Lookup which creates a dynamic alphabetically ordered lexicon of the cases having that name somewhere within the citation.

Relevance ranking can be significantly improved by weighing individual items (the default position being medium) as **high, low**, or **reduce relevance**. Key legal concepts which are unique to the particular legal issue can be marked as high, while concepts which appear frequently in several areas of the law can be marked as low or given a reduction of weight function. The present Flexicon system gives each quadrant a particular weight. These weights were formulated after extensive testing on eighteen most promising algorithms.

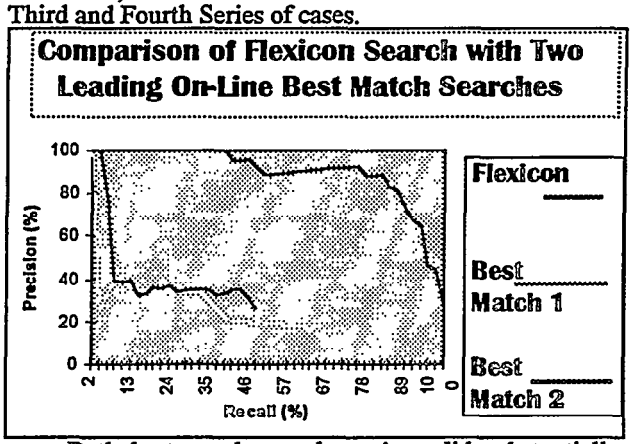
10. Evaluating Effectiveness

Retrieval effectiveness or retrieval performance, is measured in terms of a comparison between how many of the relevant documents are returned, (recall) and the proportion of relevant documents to non-relevant documents (precision) at any point of recall. [Turtle, 1995] Relevancy is measured in terms of the congruency between the information need and the returned documents. There must be a close correspondence between the nature and structure of the search engine, the representation of the document, and the nature and structure of the search query. The closer the search query corresponds to the information need, and to the representation of the text of the document, the more efficient the search engine will be.

The particular profile of the above legal problem used in the Flexicon search numbers approximately seventy terms. The first nineteen cases returned were all relevant to the

legal problem. After the first twenty cases a number of cases began to appear which dealt with issues of occupiers liability but the harmful act was the result of negligence rather than intent. We considered these not to be relevant as the legal issue is somewhat different. Performance is often portrayed by use of a graph with the vertical axis showing precision and the horizontal axis showing recall.

The graph below compares the performance of Flexicon on the above problem with that of two leading best-match search engines using the above natural language query, formulated by an experienced legal researcher, which was then used on the same California Third and Fourth Series of cases.



Both best-match search engines did substantially better than the Boolean search, and performed very close to each other so far as the results were concerned. In all cases the test of relevancy was the same, conformity with the factual script. Semi-relevant cases were treated as non-relevant. Only cases which involved the failure of an occupier of land or premises to prevent a wrongful act of a third party to an invitee or a licensee on the property itself, was considered to be relevant.

The precision for the first nineteen documents returned by our profile search is 100%, and then it dips slightly with the first non-relevant case, it gradually drops slightly more up to the fortieth case, but then drops substantially. This drop indicates that a substantial number of non-relevant cases are now being returned. The straighter the line, the more efficient is the system. A system which returns more relevant documents than another system and returns a higher number of documents near the top of a ranking is considered to perform better than one which retrieves a smaller number of cases, and with less of the relevant documents at the top of the ranking. A perfect performance would be to return all of the relevant documents first, then the semi-relevant documents, where there is a minor modification from the factual script, to be then followed by the first non-relevant document. There is another element, however, which must be taken into account, particularly where recall is concerned. The lower the precision in terms of increasing numbers of non-relevant decisions, the greater the amount of time it takes to ascertain percentage of recall. Theoretically one can, in fact, recover all of the decisions in a database, or a set of case reports, if one is willing to spend that amount of time.

The above Boolean query retrieved four cases, all on point out of 46. The two best match search engines, using the above natural language query recovered approximately half of the relevant cases. It does not follow from this that best match outperforms exact match. It depends upon the nature of the information need. Equally, the implications one can draw as to the comparison between the Flexicon search engine and best match are limited, as it was only done on a single problem. This kind of testing is extremely time consuming as it takes a great deal of research to find all of the relevant cases in a large legal database.

The difficulty of information retrieval can be briefly stated, but the solution is far from simple. Text can be divided up by the computer into single words. Humans think, speak, and write substantially in terms of multiple word groupings. It takes a good deal of intelligence to recognize multiple word concepts. The Flexicon process is designed to facilitate the use of multiple word concepts by permitting the user to create dynamic multiple word lexicons around single key words and to create an information -profile in terms of a factual script where the occupier, the nature of the premises, the location where the wrongful act took place, the status of the victim, the status of the third party doing the wrong, the character of the harm, and the nature of the security measure failure, are the significant variables.

11 Lexicons and Legal Reasoning

Even though Deep Blue has defeated Garry Kasparov, I still personally believe that the human race will never celebrate the birthday of HAL or any kind of artificial intelligence system with the capabilities imagined by Clarke and Kubrick in *2001: A Space Odyssey* nor do I believe, unlike some, [D'Amato] that a computer will ever be able to replace a judge. Time, however, will tell. The problem with legal reasoning lies in the fact that judicial decision making requires the judge to apply legal doctrine to sets of facts, and the associative relationships between doctrine and facts cannot be formalized. The most relevant case returned on the above set of facts was *Nola M. v. University of Southern California*, 20 Cal. Rptr. 2d. 97, having very similar facts. The headnote summarizes the doctrinal reasoning as follows:

> *For the actor to be liable... there has to be a duty and a breach of that duty...*
>
> *Duty is question of law to be determined on case-by-case basis.*
>
> *If court finds defendant was under duty to protect plaintiff, trier of fact must then decide whether defendant's protective measures were reasonable under the circumstances, that is, whether there was breach of defendant's duty of care.*
>
> *When no duty of landowner to protect another exists... it does not matter where land is located or who has previously done what to whom... nor... how many invitees have been maimed or murdered.*

The above ratio is, of course, completely circular. No matter how many cases or texts one reads on the above problem, it will be impossible to come up with a formalized set of rules which will link the legal doctrine to factual situations.

What the Flexicon system permits one to do is to retrieve almost all of the relevant cases in a large database. One can then separate the ones where liability was found and isolate the teleological considerations which underlie the relationships between the legal doctrine of occupier's liability and the factual situations. For example by analyzing the cases where liability was found in the 45

relevant cases returned by the Flexicon search, it becomes very clear that there were three factual elements present in almost all of the cases where liability was found: 1) the occupier was in full possession and control of the premises, 2) there were prior similar incidents of which the occupier had knowledge, and 3) the cost of the security measure was relatively low. The Flexicon system would permit lawyers to a) search large databases of cases, b) separate the cases where liability is found, c) isolate the factors which those cases have in common, and d) give an explanation for the decision in terms of the underlying teleological structure of the law which furnishes the associative relationships between legal doctrine and the facts.

Rather than viewing the computer and the human as two different intelligent systems struggling to communicate with each other, a more useful way to view the relationship between the human and the machine is to view the machine as an extension of the human information system. For the time, being, at least, our time and our resources probably can be more profitably spent by seeking methods of solving *difficult tasks* through the process of *creative transformations* than by seeking to replicate or simulate human intelligence in the machine.

12 Conclusion

It has been one of the fundamental objectives of the FLAIR Project to demonstrate the importance of integrating legal theory and psycho-linguistics with artificial intelligence and law. Simplistic paradigms of human language, legal reasoning, and legal discourse block progress in artificial intelligence and law. The fetishization of logic which historically has permeated the field of artificial intelligence research, while understandable in terms of the underlying structure of the computer, doesn't fit the way humans think. Lexicons, on the other hand, are a common denominator for computers, books, and humans. Computational lexicology offers an alternative and more promising direction for research in the field of artificial intelligence and law.

The lexical structure of the Flexicon information system is shown in the diagram below. Legal case data is put through a database building process, in conjunction with our system lexicons and recognition tools. Flexicon has three system lexicons: a hierarchical legal concept lexicon, and alphabetical noise word and first name lexicons. The completed database building process results in eight database lexicons; namely, five dynamic, alphabetical lexicons (legal concepts, cited cases, cited statutes, proper names, and facts), two static, hierarchical lexicons (legal concepts, and cited statutes), and one static, alphabetical lexicon (cases in database). The database lexicons are then combined into the final product, a Flexicon database.

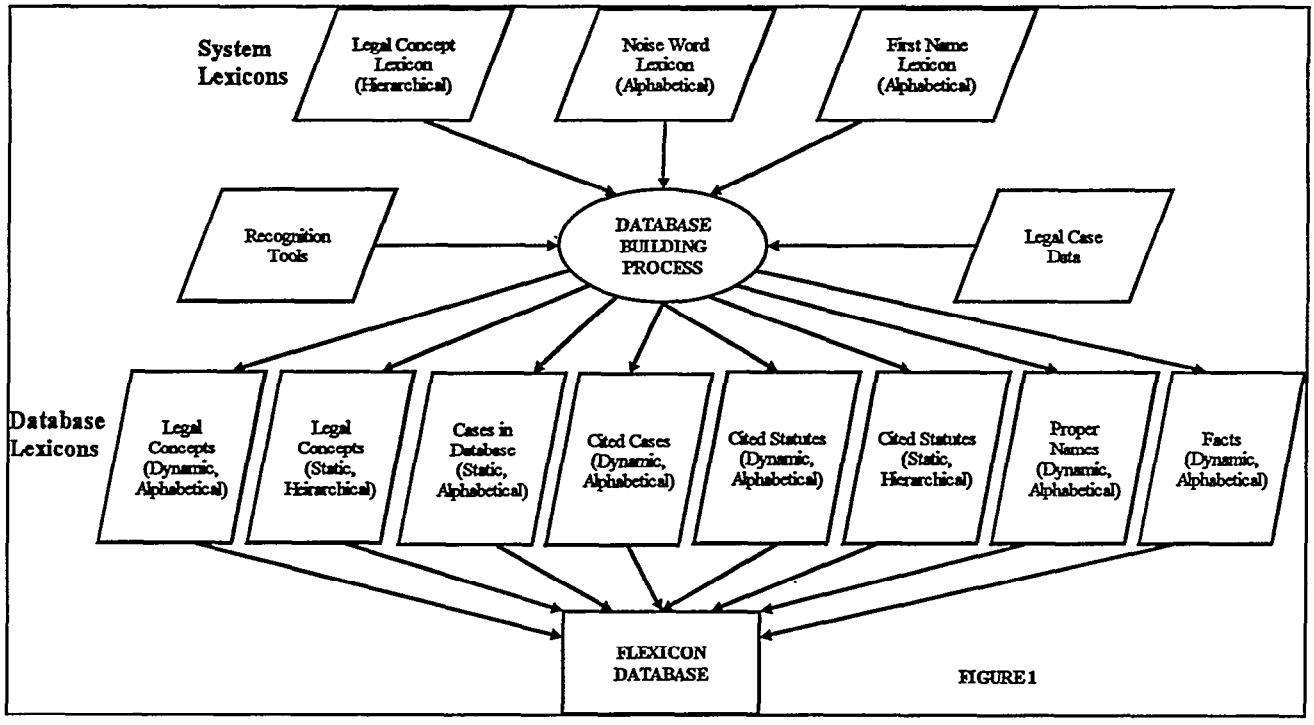


FIGURE1

Reducing documents and large bodies of text to a variety of word lists or lexicons is a form of simplified text representation. This permits the computer to perform a variety of simplified computational processes which, after inverse transformations, completes tasks which would otherwise be extremely difficult. The Flexicon system demonstrates that lexical structures can substantially improve the quality and precision of information retrieval in

large legal databases. In particular, the use of a lexical design permits a high degree of relevance ranking with a full multiple word item facility. It enables one to form a much more detailed profile of the information need and a superior document representation and profile — document matching. In addition it permits a weighting of both individual terms and categories of terms which expedites retrieval effectiveness. It provides a form of document summary (the Flexnote), and it facilitates the automatic addition of hypertext links from the quadrant document representation to the full text. It provide the means for making the content of the database more visible, and in a book-like form which is familiar to the legal researcher. Thus psycholinguistics, and the emerging field of computational lexicology has, much to offer the field of artificial intelligence and law.

References

Atkins, B. T.S. & Zampolli, eds. 1994, *Computational Approaches to the Lexicon*, Oxford U. Press.

Beckwith, R., Felbaum, C, Gross, D, & Miller, G, 1991; WordNet: a lexical database organized on psycholinguistic principles, in *Lexical acquisition: Exploiting online resources to build a lexicon*, Zernick, U ed., Erlbaum.

Boguraev, Bran & Briscoe, Ted, eds. 1989, *Computational Lexicography for Natural Language Processing*, Longman

Coval, SC, & Smith, JC, 1982, Rights, Goals, and Hard Cases, 1 *Law and Philosophy*, 451.

Coval, S. C. & Smith, JC, 1986, *Law and its Presuppositions*, Routledge and Kegan Paul Ltd., London.

Croft, W.B., Turtle, H. R., & Lewis D. D., 1991, The Use of Phrases and Structured Queries in Information Retrieval, Proceedings of *The Fourteenth International Conference on Research and Development in Information Retrieval*, ACM.

D' Amato, Anthony, 1977, Can/Should Computers Replace Judges, 11 *Georgia Law Review*, 1277.

Deedman, Cal, 1994, *Developing Conceptual Frameworks for Structuring Legal Knowledge to Build Knowledge-Based Systems*, Ph.D. thesis, UBC.

Deedman, Cal, & Smith, JC, 1991, The Nervous Shock Adviser: A Legal Expert System in Case-based Law, *Operational Expert Systems Applications in Canada*, Ching Y. Suen and Rajjan Shinghai eds., Pergamon Press Oxford, 56.

Derrida, Jacques, 1976, *Of Grammatology*, John Hopkins U. Press, Baltimore.

Derrida, Jacques, 1978, *Writing and Difference*, Routledge, London.

Garfinkel, Simon, (1997) Happy Birthday Hal, *Wired*, 120, (January).

Guo, Cheng-Ming, 1995, *Machine Tractable Dictionaries*, Ablex.

Lacan, Jacques, 1977, *Écrits*, New York, W, W. Norton.

Lacan, Jacques, 1988, *The Seminar of*, Book II, New York, W, W. Norton.

Lacan, Jacques, 1993, *The Seminar of*, Book III, New York, W, W. Norton.

Kowalski, Andrzej, 1991, Case-Based Reasoning and the Deep Structure Approach to Knowledge Representation, Proceedings of *The Third International Conference on International Intelligence on Law*, A.C.M. Press, New York, 21.

MacCrimmon, Marilyn, 1989, Expert Systems in Case-Based Law: The Hearsay Advisor, Proceedings of *The Second International Conference on International Intelligence on Law*, A.C.M. Press, New York, 68.

Melzak, Z. A., 1983, *Bypasses: A Simple Approach to Complexity*, Wiley.

de Saussure, Ferdinand, 1959, *Course in General Linguistics*, Philosophical Library, New York.

Smith, JC, 1976, *Legal Obligation*, U. of London, the Athlone Press, London.

Smith, JC, 1993, Action Theory and Legal Reasoning, *Tort Theory*, (Cooper-Stephenson, K. D., & Gibson E. eds.) Captus U. Pub.

Smith, JC & Deedman,Cal, 1987, The Application of Expert-System Technology to Case-Based Law, The Proceedings of *The First International Conference on Artificial Intelligence and Law*, A.C.M. Press, New York, 84.

Smith, JC & Gelbart, D., 1990, Towards a Comprehensive Legal Information Retrieval System, Proceedings of *The First International Conference on Database and Expert Systems Applications*, (DEXA), 121.

Smith, JC, & Gelbart, D., 1991, Beyond Boolean Search: FLEXICON, a Legal Text-Based Intelligent System, *Proceedings of the Third International Conference on Artificial Intelligence and Law*, A.C.M. Press, New York, 225.

Smith, JC, & Gelbart, D. 1993, FLEXICON: An Evolution of a Statistical Ranking Model Adopted for Intelligent Legal Text Management, Proceedings of *The Fourth International Conference on International Intelligence on Law*, A.C.M. Press, New York, 142.

Smith, JC. & Gelbart D., et al, 1995, Artificial Intelligence and Legal Discourse, 3 *Artificial Intelligence and Law*, 55.

Stork, D. G. (ed.) 1997, HAL's Legacy: 20001's Computer as Dream and Reality, MIT Press.

Turtle, Howard, 1995, Text Retrieval in the Legal World, 3 *Artificial Intelligence and Law*, 5.

Walker, D. & Zampolli, A., 1989, Forward, *Computational Lexicography for Natural Language Processing*, Boguraev, B. & Briscoe, T. (eds.) Longman.

Wilks, Yorick A., Slator, Brian M, & Guthrie, Louise M., 1996, *Electric Words*, Bradford.

Winograd, Terry, 1990, Thinking Machines: Can there be? Are we?, *The Foundations of Artificial Intelligence*,167 Derek Partridge & Yorick Wilks,(eds.,).