Foundations and Trends® in
Web Science
Vol. 1, Nos. 3–4 (2006) 199–327

DOI: 10.1561/1800000002

now
the essence of knowledge

Natural Language Processing as a Foundation of the Semantic Web

By Yorick Wilks and Christopher Brewster

Contents

Foundations and Trends® in
Web Science
Vol. 1, Nos. 3–4 (2006) 199–327

DOI: 10.1561/1800000002

now
the essence of knowledge

Natural Language Processing as a Foundation of the Semantic Web

Yorick Wilks[1] and Christopher Brewster[2]

[1] *University of Oxford, UK*
[2] *Aston University, UK, c.a.brewster@aston.ac.uk*

Abstract

The main argument of this paper is that Natural Language Processing (NLP) does, and will continue to, underlie the Semantic Web (SW), including its initial construction from unstructured sources like the World Wide Web (WWW), whether its advocates realise this or not. Chiefly, we argue, such NLP activity is the only way up to a defensible notion of meaning at conceptual levels (in the original SW diagram) based on lower level empirical computations over usage. Our aim is definitely not to claim logic-bad, NLP-good in any simple-minded way, but to argue that the SW will be a fascinating interaction of these two methodologies, again like the WWW (which has been basically a field for statistical NLP research) but with deeper content. Only NLP technologies (and chiefly information extraction) will be able to provide the requisite RDF knowledge stores for the SW from existing unstructured text databases in the WWW, and in the vast quantities needed. There is no alternative at this point, since a wholly or mostly hand-crafted SW is also unthinkable, as is a SW built from scratch

and without reference to the WWW. We also assume that, whatever the limitations on current SW representational power we have drawn attention to here, the SW will continue to grow in a distributed manner so as to serve the needs of scientists, even if it is not perfect. The WWW has already shown how an imperfect artefact can become indispensable.

1

Introduction

> *In the middle of a cloudy thing is another cloudy thing, and within that another cloudy thing, inside which is yet another cloudy thing ... and in that is yet another cloudy thing, inside which is something* ***perfectly clear and definite."***
>
> — Ancient Sufi saying

The newly developing field of Web Science has been defined as "the science of decentralised information systems" [10] which clearly covers a very broad area. Nonetheless the core focus of Web Science is the Semantic Web (SW) conceived of as a more powerful, more functional and more capable version of our current document and language centric World Wide Web (WWW). This paper focusses on the question of what kind of object this SW is to be. Our particular focus will be its semantics and the relationship between knowledge representations and natural language, a relationship concerning which this paper wishes to express a definite perspective. This is a vast, and possibly ill-formed issue, but the SW is no longer simply an aspiration in a magazine article [11] but a serious research subject on both sides of the Atlantic

and beyond, with its own conferences and journals. So, even though it may be beginning to exist in a demonstrable form, in the way the WWW itself plainly does exist, it is a topic for research and about which fundamental questions can be asked, as to its representations, their meanings and their groundings, if any.

The position adopted here is that the concept of the SW has two distinct origins, and this persists now in two differing lines of SW research: one, closely allied to notions of documents and natural language (NL) and one not. These differences of emphasis or content in the SW carry with them quite different commitments on what it is to interpret a knowledge representation and what the method of interpretation has to do with meaning in natural language.

We shall attempt to explore both these strands here, but our sympathies will be with the NL branch of the bifurcation above, a view that assumes that NL is, in some clear sense, our primary method of conveying meaning and that other methods of conveying meaning (formalisms, science, mathematics, codes, etc.) are parasitic upon it. This is not a novel view: it was once associated firmly with the philosophy of Wittgenstein [197], who we shall claim is slightly more relevant to these issues than is implied by Hirst's immortal, and satirical, line that "The solution to any problem in artificial intelligence (AI) may be found in the writings of Wittgenstein, though the details of the implementation are sometimes rather sketchy [79]."

Later parts of the paper will explore the general issue of language processing and its relevance to current, and possibly future, techniques of web searching, and we shall do this by means of an examination of the influential claims of Karen Spärck Jones that there cannot be "meaning codings" in the SW or Internet search. Having, we believe, countered her arguments, we go on to examine in particular the meaning codings expressed in ontologies, as they play so crucial a role in the SW. Our core argument is that such representations can be sound as long as they are empirically based. Finally, we turn to a methodology for giving such an empirical base to ontologies and discuss how far that program has yet succeeded.

There are a number of NLP technologies which will not be discussed here; some have relationships to the Internet but they are not yet basic

technologies in the way those of content representation and search are that we will discuss in the body of this paper. These include automatic summarisation, text mining, and machine translation (MT).

MT is the oldest of these technologies and we will touch on its role as a driver behind the introduction of statistical and data-based methods into NLP in the late eighties. MT has a history almost fifty years long, and basic descriptions and histories of its methods can be found in Nirenburg et al. [130]. The oldest functioning MT system SYSTRAN is still alive and well and is believed to the basis of many of the language translations offered on the Internet such as Babelfish MT. This is a service that translates a webpage on demand for a user with a fair degree of accuracy. The technology is currently shifting with older language pairs being translated by SYSTRAN and newer ones by empirical application of statistical methods to text corpora.

Text mining (TM) [90] is a technique that shares with Information Retrieval (IR) a statistical methodology but, being linked directly to the structure of databases, does not have the ability to develop in the way IR has in recent decades by developing hybrid techniques with NLP aspects. TM can be seen as a fusion of two techniques: first, the gathering of information from text by some form of statistical pattern learning and, secondly, the insertion of such structured data into a database so as to carry out a search for patterns within the structured data, hopefully novel patterns not intuitively observable.

Another well-defined NLP task is automatic summarisation, described in detail in [110] and which takes an information source, extracts content from it, and presents the most important content to the user in a condensed form and in a manner sensitive to the user's or application's needs. Computers have been producing summaries since the original work of [108]. Since then several methods and theories have been applied including the use of $tf * idf$ measures, sentence position, cue and title words; partial understanding using conceptual structures; cohesive properties of texts (such as lexical chains) or rhetorical structure theory (RST). Most summarisation solutions today rely on a 'sentence extraction' strategy where sentences are selected from a source document according to some criterion and presented to the user by concatenating them in their original document order. This is a robust and

sometimes useful approach, but it does not guarantee the production of a coherent and cohesive summary. Recent research has addressed the problems of sentence extracts by incorporating some NL generation techniques, but this is still in the research agenda.

We will give prominence to one particular NLP technology, in our discussion of language, the SW and the Internet itself: namely, the automatic induction of ontology structures. This is a deliberate choice, because that technology seeks to link the distributional properties of words in texts directly to the organising role of word-like terms in knowledge bases, such as ontologies. If this can be done, or even partially done, then it provides an empirical, procedural, way of linking real words to abstract terms, whose meanings in logic formulas and semantic representations has always been a focus of critical attention: how, people have always asked, can these things that look like words function as special abstract bearers of meaning in science and outside normal contexts? Empirical derivation of such ontologies from texts can give an answer to that question by grounding abstract use in concrete use, which is close to what Wittgenstein meant when he wrote of the need to "bring back words from their metaphysical to their everyday uses" [197, Section 116].

As noted above, Web Science has been defined as "the science of decentralised information systems" [10] and has been largely envisaged as the SW which is " a vision of extending and adding value to the Web, ... intended to exploit the possibilities of logical assertion over linked relational data to allow the automation of much information processing." Such a view makes a number of key assumptions, assumptions which logically underlie such a statement. They include the following:

- that a suitable logical representational language will be found;
- that there will be large quantities of formally structured relational data;
- that it is possible to make logical assertions i.e., inferences over this data consistently;
- that a sufficient body of knowledge can be represented in the representational language to make the effort worthwhile.

We will seek to directly and indirectly challenge some of these assumptions. We would argue that the fundamental decentralised information on the web is text (unstructured data, as it is sometimes referred to) and this ever growing body of text needs to be a fundamental source of information for the SW if it is to succeed. This perspective places NLP and its associated techniques like Information Extraction at the core of the Semantic Web/Web Science enterprise. A number of conclusions follow from this which we will be exploring in part in this paper.

Of fundamental relevance to our perspective is that the SW as a programme of research and technology development has taken on the mantel of artificial intelligence. When Berners-Lee stated that the SW "will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users" [11], this implied knowledge representation, logic and ontologies, and as such is a programme almost identical to which AI set itself from the early days (as a number of authors have pointed out e.g., Halpin [70]). This consequently makes the question of how NLP interacts with AI all the more vital, especially as the reality is that the World Wide Web consists largely of human readable documents.

The structure of this paper is as follows: In Section 2, we look at the SW as an inheritor of the objectives, ideals and challenges of traditional AI. Section 3 considers the competing claim that in fact the SW will consist exclusively of "trusted databases." In Section 4, we turn to the view that the SW *must* have its foundation on NL artefacts, documents, and we introduce the notion of ontology learning from text in this section as well. This is followed by a brief conclusion.

2

The SW as Good Old Fashioned AI

The relation between philosophies of meaning, such as Wittgenstein's, and classic AI (or GOFAI as it is often known: Good Old Fashioned AI [73]) is not an easy one, as the Hirst quotation cited above implies. GOFAI remains committed to some form of logical representation for the expression of meanings and inferences, even if not the standard forms of the predicate calculus. Most issues of the AI journal consist of papers within this genre.

Some have taken the initial presentation of the SW by Berners-Lee et al. to be a restatement of the GOFAI agenda in new and fashionable WWW terms [11]. In that article, the authors described a system of services, such as fixing up a doctor's appointment for an elderly relative, which would require planning and access to the databases of both the doctor's and relative's diaries and synchronising them. This kind of planning behaviour was at the heart of GOFAI, and there has been a direct transition (quite outside the discussion of the SW) from decades of work on formal knowledge representation in AI to the modern discussion of ontologies. This is clearest in work on formal ontologies representing the content of science e.g., [82, 138], where many of the same individuals have transferred discussion and research from one paradigm to the other. All this has been done within what one could call the

standard Knowledge Representation assumption within AI. This work goes back to the earliest work on systematic Knowledge Representation by McCarthy and Hayes [116], which we could take as defining core GOFAI. A key assumption of all such work was that the predicates in such representations merely look like English words but are in fact formal objects, loosely related to the corresponding English, but without its ambiguity, vagueness and ability to acquire new senses with use.

The SW incorporates aspects of the formal ontologies movement, which now can be taken to mean virtually all the content of classical AI Knowledge Representation, rather than any system of merely hierarchical relations, which is what the word "ontology" used to convey. More particularly, the SW movement envisages the (automatic) annotation of the texts of the WWW with a hierarchy of annotations up to the semantic and logical, which is a claim virtually indistinguishable from the old GOFAI assumption that the true structure of language was its underlying logical form. Fortunately, SW comes in more than one form, some of which envisage statistical techniques, as the basis of the assignment of semantic and logical annotations, but the underlying similarity to GOFAI is clear. One could also say that semantic annotation, so conceived, is the inverse of Information Extraction (IE), done not at analysis time but, ultimately, at generation time without the writer being aware of this (since one cannot write and annotate at the same time). SW is as, it were, producer, rather than consumer, IE.

It must also be noted that very few of the complex theories of knowledge representation in GOFAI actually appear within SW contributions so far. These include McCarthy and Hayes fluents, [116], McCarthy's [117] later autoepistemic logic, Hayes' Naïve Physics [74], Bobrow and Winograd's KRL [13], to name but a few prominent examples. A continuity of goals between GOFAI and the SW has not meant continuity of research traditions and this is both a gain and a loss: the gain of simpler schemes of representation which are probably computable; a loss because of the lack of sophistication in current schemes of the DAML/OIL/OWL (http://www.w3.org/TR/daml+oil-reference) family, and the problem of whether they now have the representational power for the complexity of the world, common sense or scientific.

Two other aspects of the SW link it back directly to the goals of GOFAI: one is the rediscovery of a formal semantics to "justify" the SW. This is now taken to be expressed in terms of URIs (Universal Resource Indicator: basic objects on the web), which are usually illustrated by means of entities like lists of zip codes, with much discussion of how such a notion will generalise to produce objects into which all web expressions can "bottom out." This concern, for non-linguistic objects as the ultimate reality, is of course classic GOFAI.

Secondly, one can see this from the point of view of Karen Spärck Jones's emphasis on the "primacy of words" and words standing for themselves, as it were: this aspect of the SW is exactly what Spärck Jones meant by her "AI doesn't work in a world without semantic objects" [167]. In the SW, with its notion of universal annotation of web texts into both semantic primitives of undefined status and the ultimate URIs, one can see the new form of opposition to that view, which we shall explore in more detail later in this section. The WWW was basically words if we ignore pictures, tables and diagrams for the moment but the vision of the SW is that of the words backed up by, or even replaced by, their underlying meanings expressed in some other way, including annotations and URIs. Indeed, the current SW formalism for underlying content, usually called RDF triples (for Resource Description Formalism), is one very familiar indeed to those with memories of the history of AI: namely, John-LOVES-Mary, a form reminiscent at once of semantic nets (of arcs and nodes [198]), semantic templates [183], or, after a movement of LOVES to the left, standard first-order predicate logic. Only the last of these was full GOFAI, but all sought to escape the notion of words standing simply for themselves.

There have been at least two other traditions of input to what we now call the SW, and we shall discuss one in some detail: namely, the way in which the SW concept has grown from the traditions of document annotation.

2.1 The SW Blurs the Text-Program Distinction

The view of the SW sketched above has been that the IE technologies at its base (i.e., on the classic 2001 diagram) are technologies that add

"the meaning of a text" to web content in varying degrees and forms. These also constitute a blurring of the distinction between language and knowledge representation, because the annotations are themselves forms of language, sometimes very close indeed to the language they annotate. This process at the same time blurs the distinction between programs and language itself, a distinction that has already been blurred historically from both directions, by two contrary assertions:

(1) Texts are really programs (one form of GOFAI).
(2) Programs are really texts.

As to the first, there is Hewitt's [78] claim that "language is essentially a side effect" in AI programming and knowledge manipulation. Longuet–Higgins [105] also devoted a paper to the claim that English was essentially a high-level programming language. Dijkstra's view of natural language (personal communication) was essentially that natural languages were really not up to the job they had to do, and would be better replaced by precise programs, which is close to being a form of the first view.

Opposing this is a smaller group, what one might term the Wittgensteinian opposition (2005), which is the view that natural language is and always must be the primary knowledge representation device, and all other representations, no matter what their purported precision, are in fact parasitic upon language — in the sense that they could not exist if language did not. The reverse is not true, of course, and was not for most of human history. Such representations can never be wholly divorced from language, in terms of their interpretation and use. This section is intended as a modest contribution to that tradition: but a great deal more can be found in [131].

On such a view, systematic annotations are just the most recent bridge from language to programs and logic, and it is important to remember that, not long ago, it was perfectly acceptable to assume that a knowledge representation must be derivable from an unstructured form, i.e. natural language. Thus, Woods [199] in 1975:

> "*A KR language must unambiguously represent any interpretation of a sentence (logical adequacy), have a*

> *method for translating from natural language to that representation, and must be usable for reasoning.*"

The emphasis there is a method of going from the less to the more formal, a process which inevitably imposes a relation of dependency between the two representational forms (language and logic). This gap has opened and closed in different research periods: in the original McCarthy and Hayes [116] writings on KR in AI, it is clear, as with Hewitt and Dijkstra's views (mentioned earlier), that language was thought vague and dispensable. The annotation movement associated with the SW can be seen as closing the gap in the way in which Woods described.

The separation of the annotations into metadata (as opposed to leaving them within a text, as in Latex or SGML style annotations) has strengthened the view that the original language from which the annotation was derived is dispensable, whereas the infixing of annotations in a text suggests that the whole (original plus annotations) still forms some kind of object. Notice here that the "dispensability of the text" view is not dependent on the type of representation derived, in particular to logical or quasi-logical representations. Schank [160] certainly considered the text dispensable after his conceptual dependency representations had been derived, because he believed them to contain the whole meaning of the text, implicit and explicit, even though his representations would not be considered any kind of formal KR. This is a key issue that divides opinion here: can we know that any representation whatsoever contains all and only the meaning content of a text, and what would it be like to know that?

Standard philosophical problems, like this one, may or may not vanish as we push ahead with annotations to bridge the gap from text to meaning representations, whether or not we then throw away the original text. Lewis [102] in his critique of Fodor and Katz, and of any similar non-formal semantics, would have castigated all such annotations as "markerese": his name for any mark up coding with objects still recognisably within NL, and thus not reaching to any meaning outside language. The SW movement, at least as described in this section, takes this criticism head on and continues onward, hoping URIs will

solve semantic problems because they are supposed to be dereferenced sometimes by pointing to the actual entity rather than a representation (for example to an actual telephone number).

That is to say, it accepts that a SW even if based on language via annotations, will provide sufficient "inferential traction" with which to run web-services. But is this plausible? Can all you want to know be put in RDF triples, and can they then support the subsequent reasoning required? Even when agents thus based seem to work in practice, nothing will satisfy a critic like Lewis except a web based on a firm (i.e., formal and extra-symbolic) semantics and effectively unrelated to language at all. But a century of experience with computational logic has by now shown us that this cannot be had outside narrow and complete domains, and so the SW may be the best way of showing that a non-formal semantics can work effectively, just as language itself does, and in some of the same ways.

2.2 An Information Retrieval Critique of the Semantics of the SW

Karen Spärck Jones in her critique of the SW characterised it much as we have above [170]. She used a theme she had deployed before against much non-empirical NLP stating that "words stand for themselves" and not for anything else. This claim has been the basis for successful information retrieval (IR) in the WWW and elsewhere. Content, for her, cannot be recoded in any general way, especially if it is general content as opposed to that from some very specific domain, such as medicine, where she seemed to believe technical ontologies may be possible as representations of content. As she put it mischievously: IR has gained from "decreasing ontological expressiveness".

Her position is a restatement of the traditional problem of "recoding content" by means of other words (or symbols closely related to words, such as thesauri, semantic categories, features, primitives, etc.). This task is what automated annotation attempts to do on an industrial scale. Spärck Jones' key example is (in part): "A Charles II parcel-gilt cagework cup, circa 1670." What, she asks, can be recoded there, into any other formalism, beyond the relatively trivial form: {object type: CUP}?

What, she asks, of the rest of that (perfectly real and useful) description of an artefact in an auction catalogue, can be rendered other than in the exact words of the catalogue (and of course their associated positional information in the phrase)? This is a powerful argument, even though the persuasiveness of this example may rest more than she would admit on it being one of a special class of cases. But the fact remains that content can, in general, be expressed in other words: it is what dictionaries, translations and summaries routinely do. Where she is right is that GOFAI researchers are wrong to ignore the continuity of their predicates and classifiers with the language words they clearly resemble, and often differ from only by being written in upper case (an issue discussed at length in [131]). What can be done to ameliorate this impasse?

One method is that of empirical ontology construction from corpora [19, 21], now a well-established technology, even if not yet capable of creating complete ontologies. This is a version of the Woods quote above, according to which a knowledge representation (an ontological one in this case) must be linked to some NL text to be justifiably derived. The derivation process itself can then be considered to give meaning to the conceptual classifier terms in the ontology, in a way that just writing them down *a priori* does not. An analogy here would be with grammars: when linguists wrote these down "out of their heads" they were never much use as input to programs to parse language into structures. Now that grammar rules can be effectively derived from corpora, parsers can, in their turn, produce better structures from sentences by making use of such rules in parsers.

A second method for dealing with the impasse is to return to the observation that we must take "words as they stand" (Spärck Jones). But perhaps, to adapt Orwell, not all words are equal; perhaps some are aristocrats, not democrats. On that view, what were traditionally called "semantic primitives" remain just words but are also special words: a set that form a special language for translation or coding, albeit one whose members remain ambiguous, like all language words. If there are such "privileged" words, perhaps we can have explanations, innateness (even definitions) alongside an empiricism of use. It has been known since Olney et al. [135] that counts over the words used in definitions

in actual dictionaries (Webster's Third, in his case) reveal a very clear set of primitives on which all the dictionary's definitions rest.

By the term "empiricism of use," we mean the approach that has been standard in NLP since the work of Jelinek [89] and which has effectively driven GOFAI-style approaches based on logic to the periphery of NLP. It will be remembered that Jelinek attempted to build a machine translation system at IBM based entirely on machine learning from bilingual corpora. He was not ultimately successful — in the sense that his results never beat those from the leading hand-crafted system, SYSTRAN — but he changed the direction of the field of NLP as researchers tried to reconstruct, by empirical methods, the linguistic objects on which NLP had traditionally rested: lexicons, grammars, etc. The barrier to further advances in NLP by these methods seems to be the "data sparsity" problem to which Jelinek originally drew attention, namely that language is "a system of rare events" and a complete model, at say the trigram level, for a language seems impossibly difficult to derive, and so much of any new, unseen, text corpus will always remain uncovered by such a model.

In this section, we will provide a response to Spärck Jones' views on the relationship of NLP in particular, and AI in general, to IR. In the spirit of exploring the potential identification of the SW with GOFAI (the theme of this section), we will investigate her forceful argument that AI had much to learn from that other, more statistical, discipline which has, after all, provided the basis for web search as we have it now.

We first question her arguments for that claim and try to expose the real points of difference, pointing out that certain recent developments in IR have pointed in the other direction: to the importation of a metaphor recently into IR from machine translation, which one could consider part of AI, if taken broadly, and in the sense of symbol-based reasoning that Spärck Jones intends. We go on to argue that the core of Spärck Jones' case against AI (and therefore, by extension, the SW conceived as GOFAI) is not so much the use of statistical methods, which some parts of AI (e.g., machine vision) have always shared with IR, but with the attitude to symbols, and her version of the AI claim that words are themselves and nothing else, and no other symbols can replace or code them with equivalent meanings. This claim, if true, cuts

at the heart of symbolic AI and knowledge representation in general. We present a case against her view, and also make the historical point that the importation of statistical and word-based methods into language processing, came not from IR at all, but from speech processing. IR's influence is not what Spärck Jones believes it to be, even if its claims are true.

Next, we reverse the question and look at the traditional arguments that IR needs some of what AI has to offer. We agree with Spärck Jones that many of those arguments are false but point out some areas where that need may come to be felt in the future, particularly in the use of structures queries, rather than mere strings of key words, and also because the classic IR paradigm using very long queries is being replaced by the Google paradigm of search based on two or three highly ambiguous key terms. In this new IR-lite world, it is not at all clear that IR is not in need of NLP techniques. None of this should be misconstrued as any kind of attack on Spärck Jones: we take for granted the power and cogency of her views on this subject over the years. Our aim is to see if there is any way out, and we shall argue that one such is to provide a role for NLP in web-search and by analogy in the SW in general.

2.2.1 AI and NLP in Need of IR?

Speaking of AI in the past, one sometimes refers to "classical" or "traditional" AI, and the intended contrast with the present refers to the series of shocks that paradigm suffered from connectionism and neural nets to adaptive behaviour theories. The shock was not of the new, of course, because those theories were mostly improved versions of cybernetics which had preceded classical AI and been almost entirely obliterated by it. The classical AI period was logic or symbol-based but not entirely devoid of numbers, of course, for AI theories of vision flourished in close proximity to pattern-recognition research. Although, representational theories in computer vision sometimes achieved prominence [111], nonetheless it was always, at bottom, an engineering subdiscipline with all that that entailed. But when faced with any attempt to introduce quantitative methods into classical core AI in the 1970s,

John McCarthy would always respond "But where do all these numbers come from?"

Now we know better where they come from, and nowhere have numbers been more prominent than in the field of IR, one of similar antiquity to AI, but with which it has until now rarely tangled intellectually, although on any broad definition of AI as "modelling intelligent human capacities", one might imagine that IR, like MT, would be covered; yet neither has traditionally been seen as part of AI. On second thoughts perhaps, modern IR does not fall there under that definition simply because, before computers, humans were not in practice able to carry out the kinds of large-scale searches and comparisons operations on which IR rests. And even though IR often cohabits with Library Science, which grew out of card indexing in libraries, there is perhaps no true continuity between those sub-fields, in that IR consists of operations of indexing and retrieval that humans could not carry out in normal lifetimes.

2.2.1.1 Spärck Jones' Case Against AI

If any reader is beginning to wonder why we have even raised the question of the relationship of AI to IR, it is because Spärck Jones [168], in a remarkable paper, has already done so and argued that AI has much to learn from IR. In this section, our aim is to redress that balance a little and answer her general lines of argument. Her main target is AI researchers seen as what she calls "The Guardians of content". We shall set out her views and then contest them, arguing both in her own terms, and by analogy with the case of MT in particular, that the influence is perhaps in the other direction, and that is shown both by limitations on statistical methods that MT developments have shown in the last 15 years, and by a curious reversal of terminology in IR that has taken place in the same period. However, the general purpose of this section will not be to redraw boundaries between these sub-fields, but will argue that sub-fields of NLP/AI are now increasingly hard to distinguish: not just MT, but IE and Question Answering (QA) are now beginning to form a general information processing functionality that is making many of these arguments moot. The important questions in

Spärck Jones resolve to one crucial question: what is the primitive level of language data? Her position on this is shown by the initial quotation below, after which come a set of statements based on two sources [167, 168] that capture the essence of her views on the central issues:

(1) *"One of these [simple, revolutionary IR] ideas is taking words as they stand" (1990).*
(2) *"The argument that AI is required to support the integrated information management system of the future is the heady vision of the individual user at his workstation in a whole range of activities calling on, and also creating, information objects of different sorts" (1990).*
(3) *"What might be called the intelligent library" (1990).*
(4) *"What therefore is needed to give effect to the vision is the internal provision of (hypertext) objects and links, and specifically in the strong form of an AI-type knowledge base and inference system" (1990).*
(5) *"The AI claim in its strongest form means that the knowledge base completely replaces the text base of the documents" (1990).*
(6) *"It is natural, therefore, if the system cannot be guaranteed to be able to use the knowledge base to answer questions on the documents of the form 'Does X do Y?' as opposed to questions of the form 'Are there documents about X doing Y?' to ask why we need a knowledge base" (1990).*
(7) *"The AI approach is fundamentally misconceived because it is based on the wrong general model, of IR as QA" (1990).*
(8) *"What reason can one have for supposing that the different [multimodal, our interpolation] objects involved could be systematically related via a common knowledge base, and characterised in a manner independent of ordinary language" (1990).*
(9) *"We should think therefore of having an access structure in the form of a network thrown over the underlying information objects" (1990).*

(10) *"A far more powerful AI system than any we can realistically foresee will not be able to ensure that answers it could give to questions extracted from the user's request would be appropriate" (1999).*
(11) *"Classical document retrieval thus falls in the class of AI tasks that assist the human user but cannot, by definition, replace them" (1999).*
(12) *"This [IR] style of representation is the opposite of the classical AI type and has more in common with connectionist ones" (1999).*
(13) *"The paper's case is that important tasks that can be labelled information management are fundamentally inexact" (1999).*
(14) *"Providing access to information could cover much more of AI than might be supposed" (1999).*

These quotations suffice to establish a complex position, and one should note in passing the prescience of quotations (2, 3, 4, 10) in their vision of a system of information access something like the WWW we now have. The quotations indicate four major claims in the papers from which they come, which we shall summarise as follows:

A Words are self-representing and cannot be replaced by any more primitive representation; all we, as technicians with computers, can add are sophisticated associations (including ontologies) between them (quotations 1, 10, 14).
B Core AI–KR (and hence the SW conceived as GOFAI) seeks to replace words, with their inevitable inexactness, with exact logical or at least non-word-based representations (quotations 5, 6, 9).
C Human information needs are vague: we want relevant information, not answers to questions. In any case, AI–KR (and SW-as-GOFAI) cannot answer questions (quotations 7, 8, 11, 12).
D The relationship between human reader and human author is the fundamental relationship and is mediated by relevant documents (which in SW terms refers to the level of trust).

> Anyway, systems based on association can do some kinds of (inexact) reasoning and could be used to retrieve relevant axioms in a KR system (quotations 5, 13, 14).

We should not see the issues here as simply ones of Spärck Jones's critique (based on IR) of "core," traditional or symbolic AI, for her views connect directly to an internal interface within AI itself, one about which the subject has held an internal dialogue for many years, and in many of its subareas. The issue is that of the nature of and necessity for structured symbolic representations, and their relationship to the data they claim to represent.

So, to take an example from NLP, Schank [160] always held that Conceptual Dependency (CD) representations not only represented language strings but made the original dispensable, so that, for example, there need be no access to the source string in the process of machine translation after it had been represented by CD primitives; Charniak [28] and Wilks [186] in their different ways denied this and claimed that the surface string retained essential information not present in any representation. Schank's position here can be seen as exactly the type that Spärck Jones is attacking, but it was not, of course, the only AI view.

But, more generally, the kind of AI view that Spärck Jones had in her sights proclaimed the centrality and adequacy of knowledge representations, and their independence of whatever language would be used to describe what it is in the world they represent (that is the essence of her claims A and B). The key reference for the view she rejects would be McCarthy and Hayes [116]. Its extreme opposite, in machine vision at least, would be any view that insists on the primacy of data over any representation [56]. The spirit of Chomsky, of course, hovers over the position, in language modelling at least, that asserts the primacy of a (correct) representation over any amount of data. Indeed, he produced a range of ingenious arguments as to why no amount of data could possibly produce the representations the brain has for language structure [31], and those arguments continued to echo through the dispute. For example, in the dispute between Fodor and Pollack [50, 142] as to whether or not nested representations could be derived by any form of machine learning from language data: Fodor denied this was possible

but lost the argument when Pollack [142] produced his Recursive Auto-Associative Networks that could do exactly that.

Again, and now somewhat further from core AI, one can see the issue in Schvaneveldt's Pathfinder networks [161] which he showed, in psychological experiments, could represent the expertise of fighter pilots in associationist networks of terms, a form very close to the data from which it was derived. This work was a direct challenge to the contemporary expert-systems movement for representing such expertise by means of high-level rules.

2.2.1.2 Some Countervailing Considerations from AI

It should be clear from the last paragraphs that Spärck Jones is not targeting all of AI, which might well be taken to include IR on a broad definition, but a core of AI, basically the strong representationalist tradition, one usually (but not always, as in the case of Schank above) associated with the use of first-order predicate calculus. And when one writes of a broad definition, it could only be one that does not restrict AI to the modelling of basic human functionalities, the notion behind Papert's original observation that AI could not and should not model superhuman faculties, ones that no person could have. In some sense, of course, classic IR is superhuman: there was no pre-existing human skill, as there was with seeing, talking or even chess playing that corresponded to the search through millions of words of text on the basis of indices. But if one took the view, by contrast, that theologians, lawyers and, later, literary scholars were able, albeit slowly, to search vast libraries of sources for relevant material, then on that view IR is just the optimisation of a human skill and not a superhuman activity. If one takes that view, IR is a proper part of AI, as traditionally conceived.

However, that being said, it may be too much a claim (D above) in the opposite direction to suggest, as Spärck Jones does in a remark at the end of one of the papers cited, that core AI may need IR to search among the axioms of a formalised theory [168] in order to locate relevant axioms to compose a proof. It is certain that resolution, or any related proof program, draws in potential axioms based on the

appearance of identical predicates in them (i.e., to those in the theorem to be proved). But it would be absurd to see that as a technique borrowed from or in any way indebted to IR; it is simply the obvious and only way to select those axioms that might plausibly take part in proofs.

A key claim of Spärck Jones's (in (A) and especially (B) above) is the issue one might call primitives, where one can take that to be either the predicates of a logical representation, as in McCarthy and Hayes and most AI reasoning work, or the more linguistic primitives, present in Schank's CD work and in preference semantics [49, 193]. Her argument is that words remain their own best interpretation, and cannot be replaced by some other artificial coding that adequately represents their meaning. Spärck Jones's relationship to this tradition is complex: her own thesis [166, 194] although containing what now are seen as IR clustering algorithms applied to a thesaurus, was intended, in her own words, to be a search for semantic primitives for MT. Moreover, she contributed to the definition and development of Cambridge Language Research Unit's own semantic interlingua NUDE (for "naked ideas"). That tradition has been retained in AI and computational linguistics, both as a basis for coding lexical systems (e.g., the work of Pustejovsky [145]) and as another form of information to be established on an empirical basis from corpora and can be seen in early work on the derivation of preferences from corpora by Resnik [151], Grishman and Sterling [65], Riloff and Lehnert [152] and others. Work of this type certainly involves the exploitation of semantic redundancy, both qualitatively, in the early preference work cited above, and quantitatively, in the recent tradition of work on systematic Word Sense Disambiguation which makes use of statistical methods exploiting the redundancy already coded in thesauri and dictionaries. Unless Spärck Jones really intends to claim that any method of language analysis exploiting statistics and redundancy (like those just cited) is really IR, then there is little basis to her claim that AI has a lot to learn from IR in this area, since it has its own traditions by now of statistical methodology and evaluation and, as we shall show below, these came into AI/NLP from speech research pioneered by Jelinek, and indigenous work on machine learning, and not at all from IR.

Let us now turn to another of Spärck Jones's major claims, (C above) that QA is not a real task meeting a real human need, but that the real task is the location of relevant information, presumably in the form of documents, which is IR's classic function. First, one must be clear that there has never been any suggestion in mainstream AI that its techniques could perform the core IR task. To find relevant documents, as opposed to their content, one would have to invent IR, had it not existed; there simply is no choice. IE, on the other hand, [52] is a content searching technique, usually with a representational non-statistical component, designed to access factual content directly, and that process usually assumes a prior application of IR to find relevant material to search. The application of an IR phase prior to IE in a sense confirms Spärck Jones's "primacy of relevance," but also confirms the independence and viability of QA, which is nowadays seen as an extension of IE. IE, by seeking facts of specific forms, is always implicitly asking a question (i.e., What facts are there matching the following general form?).

However, Saggion and Gaizauskas [157] has questioned this conventional temporal primacy of IR in an IE application, and has done so by pointing out that the real answers to IE/QA questions are frequently to be found very far down the (relevance based) percentiles of returns from this prior IR phase. The reason for this is that if one asks, say, "What colour is the sky?" then, in the IR phase, the term "colour/color" is a very poor index term for relevant documents likely to contain the answer. In other words, "relevance" in the IR sense (and unboosted by augmentation with actual colour names in this case) is actually a poor guide to where answers to this question are to be found, and Gaizauskas uses this point to question the conventional relationship of IR and IE/QA.

One could, at this point, perhaps reverse Spärck Jones's jibe at AI as the self-appointed "Guardians of Content" and suggest that IR may not be as much the "Guardian of relevance" as she assumes. But whatever is the case there, it seems pretty clear that wanting answers to questions is sometimes a real human need, even outside the world of TV quiz shows. The website Ask Jeeves seemed to meet some real need, even if it was not always successful, and QA has been seen as a

traditional AI task, back to the classic book by Lehnert [98]. Spärck Jones is, of course, correct that those traditional methods were not wholly successful and did not, as much early NLP did not, lead to regimes of evaluation and comparison. But that in no way reflects on the need for QA as a task.

In fact, of course, QA has now been revived as an evaluable technique (see below), as part of the general revival of empirical linguistics, and has been, as we noted already, a development of existing IE techniques, combined in some implementations with more traditional abductive reasoning [71]. The fact of its being an evaluable technique would have made it very hard for Spärck Jones later to dismiss QA as a task in the way she did earlier, since she went so far elsewhere in identifying real NLP with evaluable techniques [171].

Over a 20- year period, computational QA has moved from a wholly knowledge-based technique (as in Lehnert's work) to where it now is, as fusion of statistical and knowledge-based techniques. Most, if not all, parts of NLP have made the same transition over that period, starting with apparently straightforward tasks like part-of-speech tagging [55] and rising up to semantic and conceptual areas like word-sense disambiguation [174] and dialogue management [32] in addition to QA. We shall now return to the origin of this empirical wave in NLP and re-examine its sources, in order to support the claim that new and interesting evidence can be found there for the current relationship of AI and IR. In her paper, Spärck Jones acknowledges the empirical movement in NLP and its closeness in many ways to IR techniques, but she does not actually claim the movement as an influence from IR. We shall argue that, on the contrary, the influence on NLP that brought in the empirical revolution was principally from speech research, and in part from traditional statistical AI (i.e., machine learning) but in no way from IR. On the contrary, the influences detectable are all on IR from outside.

2.2.1.3 Jelinek's Revolution in MT and Its Relevance

A piece of NLP history that may not be familiar to AI researchers is, we believe, highly relevant here. Jelinek, Brown and others at IBM

New York began to implement around 1988 a plan of research to import the statistical techniques that had been successful in Automatic Speech Processing (ASR) into NLP and into MT in particular. DARPA supported Jelinek's system CANDIDE [24, 25] at the same time as rival symbolic systems (such as PANGLOSS [128, 129]) using more traditional methods.

The originality of CANDIDE was to employ none of the normal translation resources within an MT system (e.g., grammars, lexicons, etc.) but only statistical functions trained on a very large bilingual corpus: 200 million words of the French–English parallel text from Hansard, the Canadian parliamentary proceedings. CANDIDE made use of a battery of statistical techniques that had loose relations to those used in ASR: alignment of the parallel text sentences, then of words between aligned French and English sentences, and n-gram models (or language models as they would now be called) of the two language separately, one of which was used to smooth the output. Perhaps the most remarkable achievement was that given 12 French output words so found (output sentences could not be longer than that) the generation algorithm could determine the unique best order (out of billions) for an output translation with a high degree of success. The CANDIDE team did not describe their work this way, but rather as machine learning from a corpus that, given what they called an "equation of MT" produced the most likely source sentence for any putative output sentence.

The CANDIDE results were at roughly the 50% level, of sentences translated correctly or acceptably in a test set held back from training. Given that the team had no access to what one might call "knowledge of French", this was a remarkable achievement and far higher than most MT experts would have predicted, although CANDIDE never actually beat SYSTRAN, the standard and traditional symbolic MT system that is the world's most used system. At this point (about 1990) there was a very lively debate between what was then called the rationalist and empiricist approaches to MT, and Jelinek began a new program of trying to remedy what he saw as the main fault of his system by what would now be called a "hybrid" approach, one that was never fully developed because the IBM team dispersed.

The problem Jelinek saw is best called "data sparseness": his system's methods could not improve even applied to larger corpora of any reasonable size because language events are so rare. Word trigrams tend to be 80% novel in corpora of any conceivable size, an extraordinary figure. Jelinek therefore began a hybrid program to overcome this, which was to try to develop from scratch the standard NLP resources used in MT, such as grammars and lexicons, in the hope of using them to generalise across word or structure classes, so as to combat data sparseness. So, if the system knew elephants and dogs were in a class, then it could predict a trigram [X Y ELEPHANT] from having seen the trigram [X Y DOG] or vice versa.

It was this second, unfulfilled, program of Jelinek that, more than anything else, began the empiricist wave in NLP that still continues, even though the statistical work on learning part-of-speech tags actually began earlier at Lancaster under Leech [55]. IBM bought the rights to this work and Jelinek then moved forward from alignment algorithms to grammar learning, and the rest is the historical movement we are still part of.

But it is vital to note consequences of this: first, that the influences brought to bear to create modern empirical, data-driven, NLP came from the ASR experience and machine learning algorithms, a traditional part of AI by then. They certainly did not come from IR, as Spärck Jones might have expected given what she wrote. Moreover, and this has only recently been noticed, the research metaphors have now reversed, and techniques derived from Jelinek's work have been introduced into IR under names like "MT approaches to IR" ([8], and see below) which is precisely a reversal of the direction of influence that Spärck Jones argued for.

We shall mention some of this work in the next section, but we must draw a second moral here from Jelinek's experience with CANDIDE and one that bears directly on Spärck Jones's claim that words are their own best representations (Claim A above). The empiricist program of recreating lexicons and grammars from corpora, begun by Jelinek and the topic of much NLP in the last 15 years, was started precisely because working with self-representations of words (e.g., n-grams) was inadequate because of their rarity in any possible

data: 80% of word trigrams are novel, as we noted earlier under the term "data sparseness." Higher-level representations are designed to ameliorate this effect, and that remains the case whether those representations are *a priori* (like Wordnet, LDOCE[1] or Roget's Thesaurus) or themselves derived from corpora.

Spärck Jones could reply here that she did not intend to target such work in her critique of AI, but only core AI (logic or semantics based) that eliminates words as part of a representation, rather than adds higher-level representation to the words. There can be no doubt that even very low-level representations, however obtained, when added to words can produce results that would be hard to imagine without them. A striking case is the use of part-of-speech tags (like PROPERNOUN) where, given a word sense resource structured in the way the LDOCE is, Stevenson and Wilks [174] were able to show that those part of speech tags alone can resolve large-scale word sense ambiguity (called homographs in LDOCE) at the 92% level. Given such a simple tagging, almost all word sense ambiguity is trivially resolved against that particular structured resource, a result that could not conceivably be obtained without those low-level additional representations, which are not merely the words themselves, as Spärck Jones expects.

2.2.1.4 Developments in IR

In this section, we draw attention to some developments in IR that suggest that Spärck Jones's characterisation of the relationship of IR to AI may not be altogether correct and may in some ways be the reverse of what is the case.

That reverse claim may also seem somewhat hyperbolic, in response to Spärck Jones's original paper, and in truth there may be some more general movement at work in this whole area, one more general than either the influence of AI on IR or its opposite, namely that traditional functionalities in information processing are now harder to distinguish. This degree of interpenetration of techniques is such that it may be just as plausible (as claiming directional influence, as above) to say that MT, QA, IE, IR as well as summarisation and, perhaps a range

[1] Longman Dictionary of Contemporary English.

of technologies associated with ontologies, lexicons, inference, the SW and aspects of Knowledge Management, are all becoming conflated in a science of information access. Without going into much detail, where might one look for immediate anecdotal evidence for that view?

Salton [158] initiated CLIR (Cross-language Information Retrieval) using a thesaurus and a bilingual dictionary between languages, and other forms of the technique have used Machine-Readable Bilingual Dictionaries to bridge the language gap [5], and Eurowordnet, a major NLP tool [180], was designed explicitly for CLIR. CLIR is a task rather like MT but recall is more important and it is still useful at low rates of precision, which MT is not because people tend not to accept translations with alternatives on a large scale like "They decided to have {PITCH, TAR, FISH, FISHFOOD} for dinner."

Gaizauskas and Wilks [52] describe a system of multilingual IE based on treating the templates themselves as a form of interlingua between the languages, and this is clearly a limited form of MT. Gollins and Sanderson [58] have described a form of CLIR that brings back the old MT notion of a "pivot language" to bridge between one language and another, and where pivots can be chained in a parallel or sequential manner. Latvian-English and Latvian-Russian CLIR could probably reach any EU language from Latvian via multiple CLIR pivot retrievals (of sequential CLIR based on Russian-X or English-X). This IR usage differs from MT use, where a pivot was an interlingua, not a language and was used once, never iteratively. Oh et al. [133] report using a Japanese–Korean MT system to determine terminology in unknown language. Gachot et al. [51] report using an established, possibly the most established, MT system SYSTRAN as a basis for CLIR. Wilks et al. [193] report using Machine Readable Bilingual Dictionaries to construct ontological hierarchies (for IR or IE) in one language from an existing hierarchy in another language, using redundancy to cancel noise between the languages in a manner rather like Gollins and Sanderson.

All these developments indicate some forms of influence and interaction between traditionally separate techniques, but are more suggestive of a loss of borderlines between traditional functionalities. More recently, however, usage has grown in IR of referring to any

technique related to Jelinek's IBM work as being a use of an "MT algorithm": this usage extends from the use of n-gram models under the name of "language models" [39, 143], a usage that comes from speech research, to any use in IR of a technique like sentence alignment that was pioneered by the IBM MT work. An extended metaphor is at work here, one where IR is described as MT since it involves the retrieval of one string by means of another [8]. IR classically meant the retrieval of documents by queries, but the string-to-string version notion has now been extended by IR researchers who have moved on to QA work where they describe an answer as a "translation" of its question [8]. On this view, QA are like two "languages." In practice, this approach meant taking FAQ questions and their corresponding answers as training pairs.

The theoretical underpinning of all this research is the matching of language models i.e. what is the most likely query given this answer, a question posed by analogy with Jelinek's "basic function of MT" that yielded the most probable source text given the translation. This can sound improbable, but is actually the same directionality as scientific inference, namely that of proving the data from the theory [16]. This is true even though in the phase where the theory is "discovered" one infers the theory from the data by a form of abduction.

2.2.1.5 Pessimism Vs Pragmatism About the Prospects of Automation

What point have we reached so far in our discussion? We have not detected influence of IR on AI/NLP, as Spärck Jones predicted, but rather an intermingling of methodologies and the dissolution of borderlines between long-treasured application areas, like MT, IR, IE, QA, etc. One can also discern a reverse move of MT/AI metaphors into IR itself, which is the opposite direction of influence to that advocated by Spärck Jones in her paper. Moreover, the statistical methodology of Jelinek's CANDIDE did revolutionise NLP, but that was an influence on NLP from speech research and its undoubted successes, not IR. The pure statistical methodology of CANDIDE was not in the end successful in its own terms, because it always failed to beat symbolic

systems like SYSTRAN in open competition. What CANDIDE did, though, was to suggest a methodology by which data sparseness might be reduced by the recapitulation of symbolic entities (e.g., grammars, lexicons, semantic annotations, etc.) in statistical, or rather machine learning, terms, a story not yet at an end. But that methodology did not come from IR, which had always tended to reject the need for such symbolic structures, however obtained — or, as one might put this, IR people expect such structures to help but rarely find evidence that they do — e.g., in the ongoing, but basically negative, debate on whether or not Wordnet or any similar thesaurus, can improve IR.

Even if we continue to place the SW in the GOFAI tradition, it is not obvious that it needs any of the systematic justifications on offer, from formal logic to URIs, to annotations to URIs: it may all turn out to be a practical matter of this huge structure providing a range of practical benefits to people wanting information. Critics like Nelson [127] still claim that the WWW is ill-founded and cannot benefit users, but all the practical evidence shows the reverse. Semantic annotation efforts are widespread, even outside the SW, and one might even cite work by Jelinek and Chelba [29], who are investigating systematic annotation to reduce the data sparseness that limited the effectiveness of his original statistical efforts at MT. One could generalise here by saying that symbolic investigations are intended to mitigate the sparseness of data (but see also Section 4.2 on another approach to data sparseness).

Spärck Jones's position has been that words are just themselves, and we should not become confused (in seeking contentful information with the aid of computers) by notions like semantic objects, no matter what form they come in, formal, capitalised primitives or whatever. However, this does draw a firm line where there is not one: we have argued in many places — most recently against Nirenburg in [131] — that the symbols used in knowledge representations, ontologies, etc., throughout the history of AI, have always appeared to be English words, often capitalised, and indeed are, in spite of the protests of their users, no more or less than English words. If anything else, they are slightly privileged English words, in that they are not drawn randomly from the whole vocabulary of the language. Knowledge representations, annotations, etc. work as well as they do and they do with such privileged

words, and the history of machine translation using such notions as interlinguas is the clearest proof of that (e.g., the Japanese Fujitsu system in 1990). It is possible to treat some words as more primitive than others and to obtain some benefits of data compression thereby, but these privileged entities do not thereby cease to be words, and are thus at risk, like all words of ambiguity and extension of sense. In Nirenburg and Wilks [131] that was Wilks's key point of disagreement with his co-author Nirenburg who holds the same position as Carnap who began this line of constructivism in 1936 with *Der Logische Aufbau der Welt*, namely that words can have their meanings in formal systems controlled by fiat. We believe this is profoundly untrue and one of the major fissures below the structure of formal AI.

This observation bears on Spärck Jones's view of words in the following way: her position could be characterised as a democracy of words, all words are words from the point of view of their information status, however else they may differ. To this we would oppose the view above, that there is a natural aristocracy of words, those that are natural candidates for primitives in virtually all annotation systems e.g. animate, human, exist and cause. The position of this paper is not as far from Spärck Jones's as appeared at the outset, thanks to a shared opposition to those in AI who believe that things-like-words-in-formal-codings are no longer words.

Spärck Jones's position in the sources quoted remains basically pessimistic about any fully automated information process; this is seen most clearly in her belief that humans cannot be removed from the information process. There is a striking similarity between that and her former colleague Martin Kay's famous paper on human-aided machine translation and its inevitability, given the poor prospects for pure MT. We believe his pessimism was premature and that history has shown that simple MT has a clear and useful role if users adapt their expectations to what is available, and we hope the same will prove true in the topics covered so far in this paper.

2.2.2 Can Representations Improve Statistical Techniques?

We now turn to the hard question as to whether or not the representational techniques, familiar in AI and NLP as both resources and the

objects of algorithms, can improve the performance of classical statistical IR. The aim is go beyond the minimal satisfaction given by the immortal phrase about IR usually attributed to Croft "For any technique there is a collection where it will help." This is a complex issue and one quite independent of the theme that IR methods should play a larger role than they do in NLP and AI. Spärck Jones has taken a number of positions on this issue, from the agnostic to the mildly sceptical.

AI, or at least non-Connectionist non-statistical AI, remains wedded to representations, their computational tractability and their explanatory power; and that normally means the representation of propositions in some more or less logical form. Classical IR, on the other hand, often characterised as a "bag of words" approach to text, consists of methods for locating document content independent of any particular explicit structure in the data. Mainstream IR is, if not dogmatically anti-representational (as are some statistical and neural net-related areas of AI and language processing), is at least not committed to any notion of representation beyond what is given by a set of index terms, or strings of index terms along with numbers themselves computed from text that may specify clusters, vectors or other derived structures.

This intellectual divide over representations and their function goes back at least to the Chomsky versus Skinner debate, which was always presented by Chomsky in terms of representationalists versus barbarians, but was in fact about simple and numerically based structures versus slightly more complex ones.

Bizarre changes of allegiance took place during later struggles over the same issue, as when IBM created the MT system (CANDIDE, see [25]), discussed earlier, based purely on text statistics and without any linguistic representations, which caused those on the representational side of the divide to cheer for the old-fashioned symbolic MT system SYSTRAN in its DARPA sponsored contests with CANDIDE, although those same researchers had spent whole careers dismissing the primitive representations that SYSTRAN contained. Nonetheless, it was symbolic and representational and therefore on their side in this more fundamental debate! In those contests SYSTRAN always prevailed over CANDIDE for texts over which neither system had been

trained, which may or may not have indirect implications for the issues under discussion here.

Winograd [195] is often credited in AI with the first NLP system firmly grounded in representations of world knowledge yet after his thesis, he effectively abandoned that assumption and embraced a form of Maturana's autopoesis doctrine [196], a biologically based anti-representationalist position that holds, roughly, that evolved creatures like us are unlikely to contain or manipulate representations. On such a view the Genetic Code is misnamed, which is a position with links back to the philosophy of Heidegger (whose philosophy Winograd began to teach at that period at Stanford in his NLP classes) as well as Wittgenstein's view that messages, representations and codes necessarily require intentionality, which is to say a sender, and the Genetic Code cannot have a sender. This insight spawned the speech act movement in linguistics and NLP, and also remains the basis of Searle's position that there cannot therefore be AI at all, as computers cannot have intentionality.

The debate within AI itself over representations, as within its philosophical and linguistic outstations, is complex and unresolved. The Connectionist/neural net movement of the 1980's brought some clarification of the issue into AI, partly because it came in both representationalist (localist) and non-representationalist (distributed) forms, which divided on precisely this issue. Matters were sometimes settled not by argument or experiment but by declarations of faith, as when Charniak said that whatever the successes of Connectionism, he did not like it because it did not give him any perspicuous representations with which to understand the phenomena of which AI treats.

Within psychology, or rather computational psychology, there have been a number of assaults on the symbolic reasoning paradigm of AI-influenced Cognitive Science, including areas such as rule-driven expertise which was an area where AI, in the form of Expert Systems, was thought to have had some practical success. In an interesting revival of classic associationist methods, Schvaneveldt [161] developed an associative network methodology — called Pathfinder networks —- for the representation of expertise, producing a network whose content is extracted directly from subjects' responses, and whose predictive

power in classic expert systems environments is therefore a direct challenge to propositional-AI notions of human expertise and reasoning.

Within the main AI symbolic tradition, as we are defining it, it was simply inconceivable that a complex cognitive task, like controlling a fighter plane in real time, on the basis of input from a range of discrete sources of information from instruments, could be other than a matter for constraints and rules over coded expertise. There was no place there for a purely associative component based on numerical strengths of association or (importantly for Pathfinder networks) on an overall statistical measure of clustering that establishes the Pathfinder network from the subject-derived data in the first place.

The Pathfinder example is highly relevant here, not only for its direct challenge to a core area of classic AI, where it felt safe, as it were, but because the clustering behind Pathfinder networks was in fact very close, formally, to the clump theory behind the early IR work such as Spärck Jones [166] and others. Schvaneveldt and his associates later applied Pathfinder networks to commercial IR after applying them to lexical resources like the LDOCE. There is thus a direct algorithmic link here between the associative methodology in IR and its application in an area that challenged AI directly in a core area. It is Schvaneveldt's results on knowledge elicitation by associative methods from groups like pilots, and the practical difference such structures make in training, that constitute their threat to propositionality here.

This is no unique example, of course: even in more classical AI one thinks of Pearl's [139] long-held advocacy of weighted networks to model beliefs, which captured (as did fuzzy logic and assorted forms of Connectionism since) the universal intuition that beliefs have strengths, and that these seem continuous in nature and not merely one of a set of discrete strengths, and that it has proved very difficult indeed to combine elegantly any system expressing those intuitions with central AI notions of logic-based machine reasoning.

2.2.3 IE as a Task and the Adaptivity Problem

We are taking IE as a paradigm of an information processing technology separate from IR; formally separate, at least, in that one returns

documents or document parts, and the other linguistic or data-base structures. IE is a technique which, although still dependent on superficial linguistic methods of text analysis, is beginning to incorporate more of the inventory of AI techniques, particularly knowledge representation and reasoning, as well as, at the same time, finding that its hand-crafted rule-driven successes can be matched by machine learning techniques using only statistical methods (see below on named entities).

IE is an automatic method for locating facts for users in electronic documents (e.g., newspaper articles, news feeds, web pages, transcripts of broadcasts, etc.) and storing them in a database for processing with techniques like data mining, or with off-the-shelf products like spreadsheets, summarisers and report generators. The historic application scenario for IE is a company that wants, say, the extraction of all ship sinkings, from public news wires in any language world-wide, and put into a single data base showing ship name, tonnage, date and place of loss, etc. Lloyds of London had performed this particular task with human readers of the world's newspapers for a hundred years before the advent of successful IE.

The key notion in IE is that of a "template": a linguistic pattern, usually a set of attribute-value pairs, with the values being text strings. The templates are normally created manually by experts to capture the structure of the facts sought in a given domain, which IE systems then apply to text corpora with the aid of extraction rules that seek fillers in the corpus, given a set of syntactic, semantic and pragmatic constraints. IE has already reached the level of success at which IR and MT (on differing measures, of course) proved commercially viable. By general agreement, the main barrier to wider use and commercialisation of IE is the relative inflexibility of its basic template concept: classic IE relies on the user having an already developed set of templates, as was the case with intelligence analysts in US Defense agencies from whose support the technology was largely developed. The intellectual and practical issue now is how to develop templates, their filler subparts (such as named entities or NEs), the rules for filling them, and associated knowledge structures, as rapidly as possible for new domains and genres.

IE as a modern language processing technology was developed largely in the US, but with strong development centres elsewhere [38, 52, 64, 80]. Over 25 systems, world wide, have participated in the DARPA-sponsored MUC and TIPSTER IE competitions, most of which have the same generic structure as shown by Hobbs [80]. Previously, unreliable tasks of identifying template fillers such as names, dates, organisations, countries, and currencies automatically often referred to as TE, or Template Element, tasks have become extremely accurate (over 95% accuracy for the best systems). These core TE tasks were initially carried out with very large numbers of hand-crafted linguistic rules.

Adaptivity in the MUC development context has meant beating the one-month period in which competing centres adapted their system to new training data sets provided by DARPA; this period therefore provides a benchmark for human-only adaptivity of IE systems. Automating this phase for new domains and genres now constitutes the central problem for the extension and acceptability of IE in the commercial world beyond the needs of the military and government sponsors who created it.

The problem is of interest in the context of this paper, because attempts to reduce the problem have almost all taken the form of introducing another area of AI techniques into IE, namely that of machine learning, and which is statistical in nature, like IR but unlike traditional core AI.

2.2.3.1 Previous Work on ML and Adaptive Methods for IE

The application of ML methods to aid the IE task goes back to work on the learning of verb preferences in the eighties by Grishman and Sterling [65] and Lehnert et al. [96], as well as early work at BBN (Bolt, Beranek and Newman, in the US) on learning to find named expressions (NEs) [12]. Many of the developments since then have been a series of extensions to the work of Lehnert and Riloff on Autoslog [152], the automatic induction of a lexicon for IE.

This tradition of work goes back to an AI notion that might be described as lexical tuning, that of adapting a lexicon automatically to

new senses in texts, a notion discussed in [191] and going back to work like Wilks [187] and Granger [60] on detecting new preferences of words in texts and interpreting novel lexical items from context and stored knowledge. These notions are important, not only for IE in general but, in particular, as it adapts to traditional AI tasks like QA.

The Autoslog lexicon development work is also described as a method of learning extraction rules from <document, filled template> pairs, that is to say the rules (and associated type constraints) that assign the fillers to template slots from text. These rules are then sufficient to fill further templates from new documents. No conventional learning algorithm was used by Riloff and Lehnert but, since then, Soderland has extended this work by using a form of Muggleton's ILP (Inductive Logic Programming) system for the task, and Cardie [26] has sought to extend it to areas like learning the determination of coreference links.

Grishman at NYU [15] and Morgan [125] at Durham have done pioneering work using user interaction and definition to define usable templates, and Riloff and Shoen [153] have attempted to use some version of user-feedback methods of IR, including user-judgements of negative and positive <document, filled template> pairings. Related methods (e.g., [104, 176], etc.) used bootstrapping to expand extraction rules from a set of seed rules, first for NEs and then for entire templates.

2.2.3.2 Supervised Template Learning

Brill-style transformation-based learning methods are one of the many ML methods in NLP to have been applied above and beyond the part-of-speech tagging origins of ML in NLP. Brill's [23] original application triggered only on POS tags; later he added the possibility of lexical triggers. Since then the method has been extended successfully to e.g. speech act determination [159] and a Brill-style template learning application was designed by Vilain [179].

A fast implementation based on the compilation of Brill-style rules to deterministic automata was developed at Mitsubishi labs [40, 154].

The quality of the transformation rules learned depends on factors such as:

- The accuracy and quantity of the training data.
- The types of pattern available in the transformation rules.
- The feature set available used in the pattern side of the transformation rules.

The accepted wisdom of the ML community is that it is very hard to predict which learning algorithm will produce optimal performance, so it is advisable to experiment with a range of algorithms running on real data. There have as yet been no systematic comparisons between these initial efforts and other conventional ML algorithms applied to learning extraction rules for IE data structures (e.g., example-based systems such as TiMBL [42] and ILP [126]. A quite separate approach has been that of Ciravegna and Wilks [36] which has concentrated on the development of interfaces (ARMADILLO and MELITA) at which a user can indicate what taggings and fact structures he wishes to learn, and then have the underlying (but unseen) system itself take over the tagging and structuring from the user, who only withdraws from the interface when the success rate has reached an acceptable level.

2.2.3.3 Unsupervised Template Learning

We should also remember the possibility of unsupervised notion of template learning: a Sheffield PhD thesis by Collier [37] developed such a notion, one that can be thought of as yet another application of the early technique of Luhn [107] to locate statistically significant words in a corpus and then use those to locate the sentences in which they occur as key sentences. This has been the basis of a range of summarisation algorithms and Collier proposed a form of it as a basis for unsupervised template induction, namely that those sentences, with corpus-significant verbs, would also contain sentences corresponding to templates, whether or not yet known as such to the user. Collier cannot be considered to have proved that such learning is effective, only that some prototype results can be obtained. This method is related, again

via Luhn's original idea, to methods of text summarisation (e.g., the British Telecom web summariser entered in DARPA summarisation competitions) which are based on locating and linking text sentences containing the most significant words in a text, a very different notion of summarisation from that discussed below, which is derived from a template rather than giving rise to it.

2.2.4 Linguistic Considerations in IR

Let us now quickly review the standard questions, some unsettled after 30 years, in the debate about the relevance of symbolic or linguistic (or AI taken broadly) considerations in the task of information retrieval.

Note too that, even in the form in which we shall discuss it, the issue is not one between high-level AI and linguistic techniques on the one hand, and IR statistical methods on the other. As we have shown, the linguistic techniques normally used in areas like IE have in general been low-level, surface orientated, pattern-matching techniques, as opposed to more traditional concerns of AI and linguistics with logical and semantic representations. So much has this been the case that linguists have in general taken no notice at all of IE, deeming it a set of heuristics almost beneath notice, and contrary to all long held principles about the necessity for general rules of wide coverage. Most IE has come from a study of special cases and rules for particular words of a language, such as those involved in template elements (countries, dates, company names, etc.).

Again, since IE has also made extensive use of statistical methods, directly and as applications of ML techniques, one cannot simply contrast statistical (in IR) with linguistic methods used in IE as Spärk Jones [169] does when discussing IR. That said, one should note that some IE systems that have performed well in MUC/TIPSTER — Sheffield's old LaSIE system would be an example [52] — did also make use of complex domain ontologies, and general rule-based parsers. Yet, in the data-driven computational linguistics movement in vogue at the moment, one much wider than IE proper, there is a goal of seeing how far complex and "intensional" phenomena of semantics and pragmatics

(e.g., dialogue pragmatics as initiated in Samuel et al. [159]) can be treated by statistical methods.

A key high-level module within IE has been co-reference, a topic that linguists might doubt could ever fully succumb to purely data-driven methods since the data is so sparse and the need for inference methods seems so clear. One can cite classic examples like:

> {*A Spanish priest*} *was charged here today with attempting to murder the Pope.* {*Juan Fernandez Krohn*}*, aged 32, was arrested after* {*a man armed with a bayonet*} *approached the Pope while he was saying prayers at Fatima on Wednesday night. According to the police,* {*Fernandez*} *told the investigators today that he trained for the past six months for the assault. He was alleged to have claimed the Pope "looked furious" on hearing* {*the priest's*} *criticism of his handling of the church's affairs. If found guilty,* {*the Spaniard*} *faces a prison sentence of 15–20 years.* (The London Times 15 May 1982, example due to Sergei Nirenburg)

This passage contains six different phrases {enclosed in curly brackets} referring to the same person, as any reader can see, but whose identity seems *a priori* to require much knowledge and inference about the ways in which individuals can be described; it is not obvious that even buying all the Google ngrams (currently for $150) from one trillion words would solve such an example (see http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html).

There are three standard techniques in terms of which this infusion (of possible NLP techniques into IR) have been discussed, and we will mention them and then add a fourth.

(i) Prior WSD (automatic word sense disambiguation) of documents by NLP techniques i.e. so that text words, or some designated subset of them, are tagged to particular senses.
(ii) The use of thesauri in IR and NLP and the major intellectual and historical link between them.

(iii) The prior analysis of queries and document indices so that their standard forms for retrieval reflect syntactic dependencies that could resolve classic ambiguities not of type (i) above.

Topic (i) is now mostly regarded as a diversion as regards our main focus of attention in this paper; even though large-scale WSD is now an established technology at the 95% accuracy level [174], there is no reason to believe it bears on this issue, largely because the methods for document relevance used by classic IR are in fact very close to some of the algorithms used for WSD as a separate task [201, 202]. IR may well not need a WSD cycle because it constitutes one as part of the retrieval process itself, certainly when using long queries as in the Text Retrieval Conference (TREC), although short web queries are a different matter, as we discussed above.

This issue has been clouded by the "one sense per discourse" claim of Yarowsky [201, 202], a claim that has been contested by Krovetz [93] who has had no difficulty showing that Yarowsky's figures (that a very high percentage of words occur in only one sense in any document) are wrong and that, outside Yarowsky's chosen world of encyclopaedia articles, it is not at all uncommon for words to appear in the same document bearing more than one sense on different occasions of use. However, if Yarowsky is taken to be making a weaker claim, about one *homonymous*-sense-per-discourse, and referring only to major senses such as between the senses of "bank", then his claim appears true.

This dispute is not one about symbolic versus statistical methods for tasks, let alone AI vs IR. It is about a prior question as to whether there is any serious issue of sense ambiguity in texts to be solved at all, and by any method. In what follows we shall assume Krovetz has the best of this argument and that the WSD problem, when it is present, cannot be solved, as Yarowsky claimed in the one-sense-per-discourse paper, by assuming that only one act of sense resolution was necessary per text. Yarowsky's claim, if true, would make it far more plausible that IR's distributional methods were adequate for resolving the sense of component words in the act of retrieving documents, because sense

ambiguity resolution would then be only at the document level, as Yarowsky's claim makes clear.

If Krovetz is right, then sense ambiguity resolution is still a local matter within a document and one cannot have confidence that any word is monosemous within a document, nor that a document-span process will resolve such ambiguity. Hence, one will have less confidence that standard IR processes resolve such terms if they are crucial to the retrieval of a document. One will expect, *a priori*, that this will be one cause of lower precision in retrieval, and the performance of web engines confirms this anecdotally in the absence of any experiments going beyond Krovetz's own. As against this, Krovetz has been unable to show that WSD as normally practised [174] makes any real difference for classic IR outcomes, and cannot explain low precision in IR. So, the source of low precision in IR may not be the lack of WSD, but instead the lexical variation in text. It may simply be the case that long IR queries (up to 200 terms) are essentially self-disambiguating, which leaves open the interesting question of what the implications of all this are for normal Web queries, of 2.6 terms on average [77], which are certainly not self-disambiguation, and to which results from classic IR may not apply.

Let us now turn to (ii), the issue of thesauri: there is less in this link in modern times, although early work in both NLP and IR made use of *a priori* hand-crafted thesauri like Roget. Though there is still distinguished work in IR using thesauri in specialised domains, beyond their established use as user-browsing tools [30], IR moved long ago towards augmenting retrieval with specialist, domain-dependent and empirically constructed thesauri, while Salton [158] early on claimed that results with and without thesauri were much the same.

NLP has rediscovered thesauri at intervals, most recently with the empirical work on word-sense disambiguation referred to above, but has remained wedded to either Roget or more hand-crafted objects like WordNet [121]. The objects that go under the term thesaurus in IR and AI/NLP are now rather different kinds of thing, although in work like [63] an established thesaurus like WordNet has been used to expand a massive lexicon for IR, again using techniques not very

different from the NLP work in expanding IE lexicons referred to earlier.

Turning now to (iii), the use of syntactic dependencies in documents, their indices and queries, we enter a large and vexed area, in which a great deal of work has been done within IR (e.g., back to Smeaton and van Rijsbergen, 1988). There is no doubt that some web search engines routinely make use of such dependencies: take a case like:

measurements of models
as opposed to
models of measurement

which might be expected to access different literatures, although the purely lexical content, or retrieval based only on single terms, might be expected to be the same. In fact, they get 363 and 326 Google hits, respectively, but the first 20 items have no common members (query performed in 2005). One might say that this case is of type (i), i.e., WSD, since the difference between them could be captured by, say, sense tagging "models" by the methods of (i), whereas in the difference between

the influence of X on Y
and (for given X and Y)
the influence of Y on X

one could not expect WSD to capture the difference, if any, if X and Y were 'climate' and 'evolution' respectively, even though these would then be quite different requests.

These are standard types of example and have been a focus of attention, both for those who believe in the role of NLP techniques in the service of IR [175], as well as those like Spärk Jones [169] who do not accept that such syntactically motivated indexing has given any concrete benefits not available by other, non-linguistic, means. Spärck Jones' paper is a contrast between what she call LMI (Linguistically Motivated Information Retrieval) and NLI (Non-Linguistically, etc.), where the former covers the sorts of efforts described in this paper and the latter more 'standard' IR approaches. In effect, this difference always comes down to one of dependencies within, for example, a

noun phrase so marked, either explicitly by syntax or by word distance windows. So, for example, to use her own principal example:

URBAN CENTRE REDEVELOPMENTS

could be structured (LMI-wise) as

REDEVELOPMENTS of [CENTRE of the sort URBAN]

or as a search for a window in full text as (NLI-wise)

[URBAN =0 CENTRE]<4 REDEVELOPMENTS

where the numbers refer to words that can intrude in a successful match.

The LMI (Linguistically Motivated Indexing) structure would presumably be imposed on a query by a parser, and therefore only implicitly by a user, while the NLI window constraints would again presumably be imposed explicitly by the user, making the search. It is clear that current web engines use both these methods — though commercial confidentiality makes it hard to be sure of the detail of this — with some of those using LMI methods derived them directly from DARPA-funded IE/IR work (e.g., NetOWL and TextWise). The job advertisements on the Google site show that the basic division of methods at the basis of this paper has little meaning for the company, which sees itself as a major consumer of LMI/NLP methods in improving its search capacities.

Spärck Jones's conclusion is one of the measured agnosticism about the core question of the need for NLP in IR: she cites cases where modest improvements have been found, and others where LMI systems' results are the same over similar terrain as NLI ones. She gives two grounds for hope to the LMIers: first, that most such results are over queries matched to abstracts, and one might argue that NLP/LMI would come into play more with access to full texts, where context effects might be on a greater scale. Secondly, she argues that some of the more negative results may have been because of the long queries supplied in TREC competitions, in that shorter, more realistic and user-derived, web queries (which, as we noted earlier, are rarely over 2.5 terms) might show a greater need for NLP. The development of Google, although proprietary (beyond the basic page rank algorithm

that is), allows one to guess that this has in fact been the case in Internet searches.

On the other hand, she offers a general remark (and we paraphrase substantially here) that IR is after all a fairly coarse task and it may be not in principle optimisable by any techniques beyond certain limits, perhaps those we have already. Here the suggestion is that other, possibly more sophisticated, techniques should seek other information access tasks and leave IR as it is. This demarcation has distant analogies to one made within the word-sense discrimination research mentioned earlier, namely that it may not be possible to push figures much above where they now are, and therefore not possible to discriminate down to the word sense level, as oppose to the cruder homograph level, where current techniques work best, on the ground that anything "finer" is a quite different kind of job, and not a purely linguistic or statistical one, but rather one for future AI.

2.2.4.1 The Use of Proposition-Like Objects as Part of Document Indexing

This is an additional notion, which, if sense can be given to it, would be a major revival of NLP techniques in aid of IR. It is an extension of the notion of (iii) above, which could be seen as an attempt to index documents by template relations, e.g., if one extracts and fills binary relation templates (X manufactures Y; X employs Y; X is located in Y) so that documents could be indexed by these facts in the hope that much more interesting searches could in principle be conducted. An example would be to find all documents which talk about any company which manufactures drug X, where this would be a much more restricted set than all those which mention drug X.

One might then go on to ask whether documents could profitably be indexed by whole scenario templates in some interlingual predicate form (for matching against parsed queries) or even by some chain of such templates, of the kind extracted as a document summary by coreference techniques [4].

Few notions are new, and the idea of applying semantic analysis to IR in some manner, so as to provide a complex structured (even

propositional) index, go back to the earliest days of IR. In the 1960s, researchers like Gardin [54], Gross [66] and Hutchins [87] developed complex structures derived from MT, from logic or "text grammar" to aid the process of providing complex contentful indices for documents, entities of the order of magnitude of modern IE templates. Of course, there was no hardware or software to perform searches based on them, though the notion of what we would now call a full text search by such patterns so as to retrieve them go back at least to Wilks [183, 184] even though no real experiments could be carried out at that time. Gardin's ideas were not implemented in any form until [7], which was also inconclusive.

Mauldin [114], within IR, implemented document search based on case-frame structures applied to queries (ones which cannot be formally distinguished from IE templates), and the indexing of texts by full, or scenario, templates appear in Pietrosanti and Graziadio [141]. The notion is surely a tempting one, and a natural extension of seeing templates as possible content summaries of the key idea in a text [4]. If a scenario template, or a chain of them, can be considered as a summary then it could equally well, one might think, be a candidate as a document index.

The problem will be, of course, as in work on text summarisation and now QA by such methods: what would cause one to believe that an a priori template could capture they key item of information in a document, at least without some separate and very convincing elicitation process that ensured that the template corresponded to some class of user needs, but this is an empirical question and one being separately evaluated by summarisation competitions [72].

Although this indexing-by-template idea is in some ways an old one, it has not been aired lately, and like so much in this area, has not been conclusively confirmed or refuted as an aid to retrieval. It may be time to revive it again with the aid of new hardware, architectures and techniques. After all, connectionism/neural nets was only an old idea revived with a new technical twist, and it had a 10 year or more run in its latest revival. What seems clear at the moment is that, in the web and Metadata world, there is an urge to revive something along the lines of "get me what we mean, not what we say" [88]. Long-serving

IR practitioners will wince at this, but to many it must seem worth a try — and can be seen in work like [6] without yet bearing much fruit, but it was very much preliminary work at SRI — since IE does have some measurable and exploitable successes to its name (especially named entity finding) and, so the bad syllogism might go, Metadata is data and IE produces data about texts, so IE can produce Metadata.

2.2.4.2 QA Within TREC

No matter what the limitation on crucial experiments so far, another place to look for evidence of the current of NLP/AI influence on IR might be the QA track within the TREC since 1999, already touched on above in connection with IRs influence on AI/NLP, or vice versa.

QA is one of the oldest and most traditional AI/NLP tasks [61, 97], but can hardly be considered solved by those structural methods. The conflation of the rival methodologies distinguished in this paper can be clearly seen in the admitted possibility, in the TREC QA competition, of providing ranked answers, which fits precisely with the continuous notion of relevance coming from IR, but is quite counterintuitive to anyone taking a common sense view of QA, on which that is impossible. It is a question master who provides a range of differently ranked answers on the classic QA TV shows, and the contestant who must make a unique choice (as opposed to re-presenting the proffered set!). That is what answering a question means; it does not mean "the height of St. Pauls is one of [12, 300, 365, 508] feet"! A typical TREC question was "Who composed Eugene Onegin?" and the expected answer was Tchiakowsky which is not a ranking matter, and listing Gorbachev, Glazunov, etc. is no help.

There were examples in the competition that brought out the methodological difference between AI/NLP on the one hand, and IR on the other, with crystal clarity: answers could be up to 250 bytes long, so if your text-derived answer was A, but wanting to submit 250 bytes of answer meant that you, inadvertently, could lengthen that answer rightwards in the text to include the form (A and B), then your answer would become wrong in the very act of conforming to

format. The anecdote is real — and no longer applies to the current form of TREC QA — but nothing could better capture the absolute difference in the basic methodology of the approaches: one could say that AI, linguistics and IR were respectively seeking propositions, sentences and byte-strings and there is no clear commensurability between the criteria for determining the three kinds of entities. More recently, Liang et al. [103] have shown that if the queries are short (a crucial condition that separates off modern democratic and Google-based IR from the classic queries of specialists) then WSD techniques do improve performance.

2.2.5 The Future of the Relation Between NLP and IR

One can make quite definite conclusions but no predictions, other than those based on hope. Of course, after 40 years, IR ought to have improved more than it has — but its overall F-measure figures are now 40% better in the TREC ad hoc track than in basic TREC1. Yet, as Spärck Jones has shown, there is no clear evidence that NLP has given more than marginal improvements to IR, which may be a permanent condition, or it maybe one that will change with a different kind of user-derived query, and Google may be one place to watch for this technology to improve strongly. It may also be worth someone in the IE/LMI tradition trying out indexing-by-scenario templates for IR, since it is, in one form or another, an idea that goes back to the earliest days of IR and NLP, but remains untested.

It is important to remember as well that there is a deep cultural division in that AI remains, in part at least, agenda driven: in that certain methods are to be shown effective. IR, like all statistical methods in NLP as well, remains more result-driven, and the clearest proof of this is that (with the honourable exception of machine translation) all evaluation regimes have been introduced in connection with statistical methods, often over strong AI/linguistics resistance.

In IE proper, one can be moderately optimistic that fuller AI techniques using ontologies, knowledge representations and inference, will come to play a stronger role as the basic pattern matching and template element finding is subject to efficient machine learning. One may

be moderately optimistic, too, that IE may be the technology vehicle with which old AI goals of adaptive, tuned, lexicons and knowledge bases can be pursued. IE may also be the only technique that will ever provide a substantial and consistent knowledge base from texts, as CYC has failed to do over twenty years (beyond its embodiment in the LCC QA system). The traditional AI/QA task, now brought within TREC, may yield to a combination of IR and IE methods and it will be a fascinating struggle.

QA may indeed re-emerge as an AI-like problem (within TREC, because it has already done so in AQUAINT) if TREC organisers shift the focus from fact lookup to answer finding, which are two different tasks. The DTO-funded (now IARPA) AQUAINT program (which must be mentioned as it is now probably the most advanced effort in QA) looks beyond fact recall and considers the user as a part of QA process. The current TREC paradigm is ill-suited for evaluating or even encouraging such research. There is an ongoing effort within AQUAINT to develop an evaluation methodology that would support user-in-the-loop evaluation in realistic tasks with the user judging his or her experience against expectations, and other users' performance. We can ask whether this is still IR, but it approximates closely to the iterative feedback processes with which real users get information from the Web, and so that must be IR in a broad sense.

The curious tale above, of the use of "translation" with IR and QA work, suggests that terms are very flexible at the moment and it may not be possible to continue to draw the traditional demarcations between IR and these close and merging NLP applications such as IE, MT and QA.

3

The SW as Trusted Databases

3.1 A Second View of the SW

There is a second view of the SW, different from the GOFAI view, one close to Berners-Lee's own vision of the SW, as expressed in Berners-Lee et al. [11], that emphasises databases as the core of the SW. In this vision, the meanings of the databases' features are kept constant and trustworthy by a cadre of guardians of their integrity, a matter quite separate from both logical representations (dear to GOFAI) and to any language-based methodology of the kind we will describe in the next section. Berners-Lee's view deserves extended discussion and consideration that cannot be given here, but it will inevitably suffer from the difficulty of any view (like GOFAI) that seeks to preserve predicates, features, facets or whatever from the NLP vagaries of sense change and drift with time. We still "dial numbers" when we phone even though that no longer means the action it did a few decades ago; hence not even number-associated concepts are safe from time. The long-running CYC project [99, 100, 101], one of the predecessors of the SW as a universal repository of formalised knowledge, suffered from precisely this difficulty of "predicate drift": that predicates did not mean this year what coders meant by them 20 years earlier. The SW has at present no solution to offer to this problem.

Berners-Lee's view has the virtues and defects of Putnam's later theory of meaning [148, 149], where scientists become the guardians of meaning, since only they know the true chemical nature of, say, molybdenum and how it differs from the phenomenally similar aluminium. Hence only these guardians know the meaning of molybdenum, independently of how it appears (which is just like aluminium!). It was essential to his theory that the scientists did not allow the criteria of meaning to leak out to the general public, lest they became subject to change. For Putnam [148], only scientists know the distinguishing criteria for water and deuterium dioxide (heavy water) which seem the same to most of the population but are not. Many observers, including one of the authors of this paper [119], have argued this separation cannot be made, in principle or in practice, since scientists are only language users in lab coats.

The idea that the SW will restrict itself merely to databases which are curated and managed is also impractical apart from reflecting a poverty of vision. The reality is that the vast majority of human knowledge is expressed and encoded in huge quantities of text. The web provides an extraordinary storehouse of information, analysis and evaluation across every conceivable topic and to ignore all that is expressed in those texts will make the SW of use only on a very limited scale and miss out of 99% of "what is there." A practical example of this phenomenon is the fact that over 2000 articles are added to the PubMed database *every day*. Only a small proportion of this information is currently processed by human beings in order to feed databases such as the numerous pathways databases. Furthermore, empirically in a number of knowledge management situations where the application of SW technologies show great potential, there do not exist the appropriate databases. For example, in an enterprise knowledge management context, where different departments are being encouraged to exchange information (e.g., between design, manufacturing and sales departments), there usually does not exist a unifying single institutional database on which to piggy-back a SW solution to their knowledge sharing problems. What *do* exist, however, are large collections of documents representing the world view and the concerns of each department.

3.2 The SW and the Representation of Tractable Scientific Knowledge

The issues concerned with Berners-Lee's "scientific database" view of the SW can be illustrated concretely by turning some to questions of meaning and interpretation of formal knowledge raised first by Kazic in connection with biological databases, which could be expected to form part of any SW wide enough to cover scientific and technical knowledge. Kazic (2006) has posed a number of issues close in spirit to those of this section, but against a background of expert knowledge of biology that is hard to capture here without more exposition than was needed in a Biocomputing proceedings, where she published. Broadly, and using arbitrary names for terms like "thymidine phosphorylase", she draws attention to two symmetric chemical reactions of "cleavage" we may write as:

$A \longleftrightarrow B$
and
$C \longleftrightarrow D$

An enzyme Z (actually EC 2.4.2.4) catalyses both reactions above according to the standard knowledge structures in the field (KEGGs maps: http:// www.genome.-ad.jp/kegg/kegg1.html). But Z is not in the class Y (a purine nucleoside) and so should not, in standard theory, be able to catalyse the two reactions above, normally the province of Y compounds, yet it does. There is a comment in the KEGG maps saying that Z can catalyse reactions like those of another enzyme Z' (EC 2.4.2.6) under some circumstances, where Z' actually is a Y, although its reactions are quite different from Z, and they cannot be substituted for each other, and neither can be rewritten as the other. Moreover, Z has apparently contradictory properties, being both a statin (which stops growth) and a growth factor. Kazic asks "so how can the same enzyme stimulate the growth of one cell and inhibit the growth of another?" (p. 2).

This is an inadequate attempt to state the biological facts in this non-specialist form, but it is clear that something very odd is going on here, something that Marxists might once have hailed as a dialectical or contradictory relationship. It is certainly an abstract structure

that challenges conventional knowledge representations, and is far more complex than the standard form of default reasoning in AI, on which view, if anything is an elephant it has four legs even though Clyde, undoubtedly an elephant, has only three.

The flavour of the phenomena here is that of extreme context dependence, that is to say, that an entity behaves quite differently — indeed in opposite fashions — in the presence of certain other entities. Languages are, of course, full of such phenomena, such as when "cleave to the Lord" and "cleave a log" mean exactly opposite things, and we have structures in language representation for describing and representing such phenomena, though there is no reason at the moment to believe they are of any assistance here.

The point Kazic is making is that it will be a requirement on any SW that represents biological information (and licences correct inferences) that it can deal with phenomena as complex as this. At first sight such phenomena seem beyond those within a standard ontology dependent on context-free relations of inclusion and the other standard relations: "To ensure the scientific validity of the SW's computations, it must sufficiently capture and use the semantics of the domain's data and computations."(p. 2) In connection with the initial translation into RDF for, she continues: "Building a tree of phrases to emulate binding ... forces one to say explicitly something one may not know (e.g. whether the binding is random or sequential, what the order of any sequential binding is ...). By expanding the detail to accommodate the phrasal structure, essential and useful ambiguities have been lost." (*ibid*)

The last quotation is revealing about the structure of science, and the degree to which it remains in parts a craft skill, even in the most technical modern areas. Even if that were not the case, being forced to be more explicit and to remove ambiguities could only be a positive influence. The quotation brings out the dilemma in some parts of advanced science that intend to make use of the SW: that of whether the science is yet explicit enough and well understood enough to be formally coded, a question quite separate from issues of whether the proposed codings (from RDF to DAML/OIL) have the representational power to express what is to be made explicit. If it is not, then Biology

may not be so different from ordinary life as we may have thought, certainly not so different from the language of auction house catalogues, in Spärck Jones' example, where the semantics remains implicit, in the sense of resting on our human interpretation of the words of annotations or comments (in this case in the margins of KEGG maps).

The analogy here is not precise, of course: the representational styles in the current SW effort have, to some degree, sacrificed representational sophistication to computational tractability (as, in a different way, the WWW itself did in the early 1990s). It may be that, when some of the greater representational powers in traditional GOFAI work are brought to bear, the KEGG-style comments may be translated from English phrases, with an implicit semantics, to the explicit semantics of ontologies and rules. It is what we must all hope for. But in the case of Spärck Jones' C17 cup description, the problem does not lie in any knowledge representation, but only in the fact that the terms involved are all so precise and specific that no generalisations — no imaginable "auction ontology" — would provide a coding that would enable the original English to be thrown away. The possibility always remains of translation into another language, or an explicit numbering of all the concepts in the passage, but there is no representational saving to be made there in either case (see here, as always, McDermott [118] for a classic demolition of that very possibility).

Kazic goes on to argue that one effect of these difficulties about explicitness is that "most of the semantics are pushed onto the applications" (p. 7), where the web agents may work or not, but there is insufficient explicitness to know why in either case. This is a traditional situation: as when a major AI objection to the connectionist/neural net movement was that, whether it worked or not, nothing was served scientifically if what it did was not understood, that is to say, transparent and explicit. There is not yet enough SW data yet to be sure, but it is completely against the spirit of the SW that its operations should be unnecessarily opaque or covert. That becomes even clearer if one sees the SW as the WWW "plus the meanings" where only additional, rather than less explicit, information would be expected.

Discussions in this area normally avoid more traditional ontological enquiry, namely what things there are in the world. Ancient questions

have a habit of returning to bite one at the end though, in this section we take a robust position, in the spirit of Quine [150] that whatever we put into our representations — concepts, sets, etc. — has existence, at least as a polite convention. But it may be that a fully explicit SW has to make ontological commitments of a more traditional sort, at least as regards the URIs: the points where the SW meets the world of unique descriptions of real things. But scientific examples of this interface in the world of genes are by no means straightforward.

Suppose we ask: what are the ontological "objects" in genetics, say in the classic Drosophila database FlyBase [124]? FlyBase ultimately grounds its gene identifiers — the formal gene names — in the sequenced Drosophila genome and associates nucleotide sequences parsed into introns, exons, regulatory regions and so on with gene ids. However, these sequences often need modifying on the basis of new discoveries in the literature: e.g., new regulatory regions "upstream" from the gene sequence are quite frequently identified, as understanding of how genes get expressed in various biological processes increases. Thus the "referent" of the gene id. changes and with it information about the role of the "gene." However, for most biologists the gene is still the organising concept around which knowledge is clustered, so they will continue to say quite happily that the gene "rutabaga" does so-and-so, even if they are aware that the referent of rutabaga has changed several times, and in significant ways, over the last decade. The curators and biologists are, for the most part, content with this, though the argument that the Drosophila community has been cavalier with gene naming has been made from within it.

This situation, assuming the non-expert description above is broadly correct, is of interest here because it shows there are still ontological issues in the original sense of that word: i.e. as to what there actually IS in the world. More precisely, it directly refutes Putnam's optimistic theory (1975, cited elsewhere in this section) that meaning can ultimately be grounded in science, because, according to him, only scientists know the true criteria for selecting the referents of terms. The Drosophila case shows this is not so, and in some cases the geneticists have no more than a hunch, sometimes proved false in practice, that there are lower level objects unambiguously corresponding to a gene

id.(and in the way SW URIs are intended to do), in the way that an elementary molecular structure, say, certainly does correspond to an element's name in Mendeleev's table.

3.3 The Need for a Third View of the SW

The importance of ontologies for the SW is undoubted, fundamentally as the means of knowledge representation but also to facilitate a number of other functions such as knowledge mapping and integration. The database view of the SW implies that ontologies will be fairly straightforward to come by. All data of relevance in the SW will be obtainable from curated databases whose database schemas clearly reflect the relevant ontology and thus obviate the need for any ontology learning [10].

Ontology development would also be straightforward if it was possible to have one overarching monolithic ontology representing the totality of human knowledge [170]. Such a world would also presuppose that all of human knowledge and understanding about the world was preordained or pre-classified, that no new subjects or research areas would be possible. Such a world would be essentially static. Alternatively, it might be argued that knowledge, and especially classification schemes, change very slowly and this can be managed by hand.

None of these arguments stands up to scrutiny. As has been discussed in this section, in the real world of empirical ontologies, knowledge is continuously changing for a number of reasons.

Ontologies correspondingly need to be seen as constructs which change constantly, reflecting the continuous alteration of our knowledge of the world. This also means that our world knowledge is uncertain, with the certainty changing as our 'knowledge' is updated.

If the SW is to have an impact in the longer term then it needs to address two fundamental issues both of which depend on the effective exploitation of NLP techniques. On the one hand, it needs to be able to link concepts/terms in ontologies with texts so as to allow, for example, semantic searches, and it will need some degree of NL understanding, some level of IE so as to enable reasoning over the data present in the texts. The significance of IE for the SW has been discussed above

in Section 2.2.2 and elsewhere [35, 36, 46]. Its importance cannot be overestimated.

On the other hand, the SW needs to apply NLP techniques to the task of ontology learning so as to provide accurate, dynamic, ontologies which reflect the current view of the world, of a domain, of a subject area, of an institution. Within the SW research program there has been a shift in discourse from *The* Semantic Web to *SWs* [112, 173] reflecting an appreciation that there needs to be multiplicity of ontologies which overlap to greater or lesser degrees for different domains. Such a growing population of ontologies can only be constructed and kept up to date by advanced language processing techniques. The application of such techniques to the SW will be discussed in the Section 4.

4

The SW Underpinned by NLP

In this section, we will explore the possibility of going beyond the GOFAI and database views of the SW in a way that is more sensitive to the vagaries of linguistic reality. If one looks at the classic SW diagram from the original Scientific American paper (see Figure 4.1), the tendency is always to look at the upper levels: rules, logic framework and proof, and it is these, and their traditional interpretations, that have caused both critics and admirers of the SW to say that it is the GOFAI project by another name. But if one looks at the lower levels one finds Namespaces and XML, which are all the products of what we may broadly call NLP obtained from the annotation of texts by a range of NLP technologies we may conveniently gather under the name IE [192].

Hence, we will argue in this section that it is entirely appropriate, indeed highly desirable, to apply NLP methods to the foundations of the SW, and that the SW will be more securely based if this happens. We will begin by rehearsing some issues of annotation at the lower end of the SW diagram. Then we shall discuss the use of NLP methods to build knowledge bases for the SW. The topic of ontologies will be introduced with a brief discussion on the possibility of the SW

being used to impose particular points of view on its users. The final, and most substantial, subsection will argue that NLP techniques will be important for ontology learning. We will begin by critiquing views opposed to the empirical development of ontologies, and then we will present the Abraxas model of ontology learning. This is both a general model of the overall parameters within which we believe ontology learning must operate and a specific instantiation of this model which generated ontologies successfully.

4.1 NL and the SW: Annotation and the Lower End of the SW Diagram

It is useful to remember that available information for science, business and everyday life, still exists overwhelmingly as text; 85% of business data still exists as unstructured data (i.e. text). So, too, of course, does the WWW, though the proportion of it that is text is almost certainly falling. And how can the WWW be absorbed into the SW except by information being extracted from natural text and stored in some other form, such as a database of facts extracted from text or annotations on text items, stored as metadata either with or separate from the texts themselves. These forms are, of course, just those provided by large-scale IE [41]. If, on the other hand, we were to take the view that the WWW will not become part of the SW, one is faced with an implausible evolutionary situation of a new structure starting up with no reference to its vast, functioning, but more primitive, predecessor. Things just do not happen like that (Figure 4.1).

XML, the annotation standard which has fragmented into a range of cognates for particular domains (e.g., TimeML, VoiceML, etc.) is the only the latest standard in a history of annotation languages. These attach codings to individual text items so as to indicate information about them, or what should be done with them in some process, such as printing. Indeed, annotation languages grew from origins as metadata for publishing documents (the Stanford roff languages, and then Knuth's Tex, later Latex), as well as semi-independently in the humanities community as a way of formalising the process of scholarly annotation of text. The Text Encoding Initiative (TEI) adopted SGML,

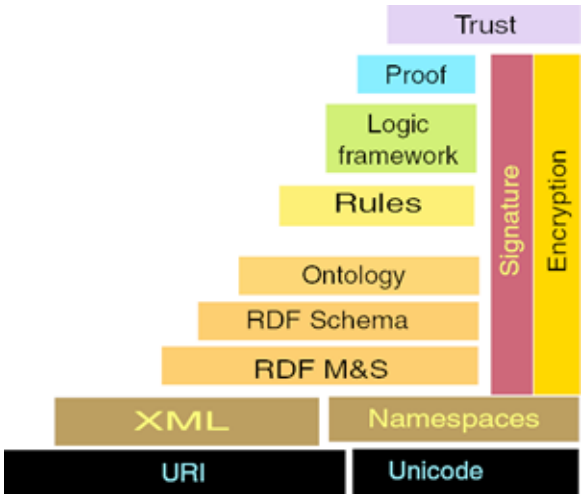

Fig. 4.1 Levels of annotation and objects in the SW (from [11]).

a development of Goldfarb's original GML.[1] SGML in turn became the origin of HTML (as a proper subset), which then gave rise to XML as well as being the genesis of the annotation movement in NLP that initially underpinned IE technology. There were early divisions over exactly how and where the annotation of text was to be stored for computational purposes; particularly between SGML, on the one hand, where annotations were infixed into the text with additional characters (as in Latex), and which had the effect of making the annotated text harder for humans to read. The DARPA research community, on the other hand, produced a functioning IE technology based on the storage of annotations (indexed by spans of characters in the text) separately as metadata, a tradition preserved in the GATE language processing platform from Sheffield [41], for example, and which now underpins many of the SW projects in Europe [14, 36]. This was one of the two origins of the metadata concept, the other being the index terms that were the basis of the standard information retrieval approach to document relevance.

[1] Recounted in Goldfarb [57].

IE is now a technology with some 25 years of history, one which began with the hand-coded approach of Sager [156] and DeJong [44], but which then moved to a fully automatic system with tools like Leech's CLAWS4 program [95] for part-of-speech tagging in 1966. This was the first program systematically to add to a text "what it meant" even at the low level of interpretation that such tags represent. IE now reliably locates names in text, their semantic types, and relates them together by means of learned structures called templates into forms of fact and events, objects virtually identical to the RDF triple stores at the basis of the SW: which are not quite logic, but very like IE output. IE began by automating annotation but now has what we may call annotation engines based on ML [36] which learn to annotate in any form and in any domain.

Extensions of this technology have led to effective QA systems trained from text corpora in well-controlled competitions and, more recently, the use of IE patterns to build ontologies directly from texts [22]. Ontologies can be thought of as conceptual knowledge structures, which organise facts derived from IE at a higher level. They are very close both to the traditional Knowledge Representation goal of AI, and occupy the middle level in the original SW diagram. We shall return to ontologies later, but for now we only want to draw attention to the obvious fact that the SW inevitably rests on some technology with the scope of IE to annotate raw texts simply to derive names, then semantic typings of entities, fact databases, and later ontologies. Where would lists of names, and nameable objects, come from if not automatically from texts; are we to imagine such inventories as simply made up by researchers?

On this view, which underlies most work on the SW and web services in Europe [132], the SW can be seen at its base level as a conversion from the WWW of texts by means of an annotation process of increasing grasp and vision, one that projects notions of meaning up the classic SW diagram from the bottom. Braithwaite [16] wrote a classic book on how scientific theories get the semantic interpretation of "high level" abstract entities (like neutrinos or bosons) from low level data; he named the process one of *semantic ascent* up a hierarchically ordered scientific theory. The view of the SW under discussion here,

which sees NLP and IE as among its foundational processes, bears a striking resemblance to that view of scientific theories in general.

4.2 The Whole Web as a Corpus and a Move to Much Larger Language Models

It may now be possible, using the whole web — and thus reducing data sparsity — to produce much larger models of a language and to come far closer to the full language model that will be needed for tasks like complete annotation and automatically generated ontologies. The Wittgensteinian will always want to look for the use rather the meaning, and nowhere has more use available than the whole web itself, even if it could not possibly be the usage of a single individual. Work will be briefly described here that seeks to make data for a language much less sparse, and without loss, by means of skip-grams. These results are as yet only suggestive and not complete, but they do seem to offer a way forward.

Kilgariff and Greffenstete [92] were among the first to point out that the web itself can now become a language corpus in principle, even though that corpus is far larger than any human could read in a lifetime as a basis for language learning. A rough computation shows that it would take about 60,000 years of constant reading for a person to read all the English documents on the WWW at the time of writing. But the issue here is not building a psychological model of an individual and so this fact about size need not deter us: Moore [122] has noted that current speech learning methods would entail that a baby could only learn to speak after a hundred years of exposure to data. But this fact has been no drawback to the development of effective speech technology — in the absence of anything better. A simple and striking demonstration of the value of treating the whole web as a corpus has been shown in experiments by e.g. Grefenstette [62] who demonstrated that the most web-frequent translation of a word pair — from among all possible translation equivalent word pairs in combination — is invariably also the correct translation.

What follows is a very brief description of the kind of results coming from the REVEAL project [69], which takes large corpora, such as

a 1.5 billion word corpus from the web, and asks how much of a test corpus is covered by the trigrams present in that large training corpus. The project considers both regular trigrams and skipgrams: which are trigrams consisting of any discontinuity of items with a maximum window of four skips between any of the members of a trigram. So, if we take the sentence:

Chelsea celebrate Premiership success.

Then the two standard trigrams in that sequence will be:
Chelsea celebrate Premiership
celebrate Premiership success
But one-skip trigrams will be:

Chelsea celebrate success
Chelsea Premiership success

which seem at least as informative, intuitively, as the original trigrams and our experiments suggest that, surprisingly, skipgrams do not buy additional coverage at the expense of producing nonsense. Recent work shows the use of skipgrams can be more effective than increasing the corpus size. In the case of a 50 million word corpus, similar results (in terms of coverage of test texts) are achieved using skipgrams as by quadrupling corpus size. This illustrates a possible use of skipgrams to expand contextual information to get something closer to 100% coverage with a (skip) trigram model, combining greater coverage with little degradation, and thus achieving something much closer to Jelinek's original goal for an empirical corpus linguistics (Figure 4.2).

The 1.5 billion word training corpus gives a 67%+ coverage by such trigrams of randomly chosen 1000 word test texts in English, which is to say 67% of the trigrams found in any random 1000 passage of English were already found in the gigaword corpus. But we obtained 74% coverage with 4 skipgrams, which suggests, by extrapolation, that it would need $75 \times 10 \times 10$ words to give 100% trigram coverage (including skipgrams up to 4 grams). Our corpus giving 74% coverage was $15 \times 10 \times 8$ words, and Greffenstette (2003) calculated there were over 10×11 words of English on the web in 2003 (i.e. about 12 times what Google indexed at that time), so the corpus needed for complete coverage of training texts by trigrams would be about seven

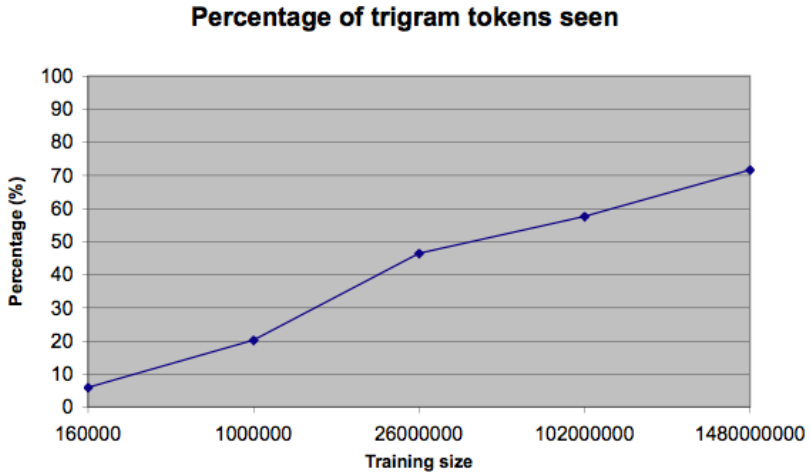

Fig. 4.2 Percentage of trigrams seen with training corpus size.

times the full English web in 2003, which is somewhat closer to the size of today's (2007) English web.

All this is, again, preliminary and tentative, but it suggests that an empiricism of usage may now be more accessible (with corpora closer to the whole web) than Jelinek thought at the time (1990) of his major MT work at IBM.

Such modern web corpora are so vast they cannot conceivably offer a model of how humans process semantics, so a cognitive semantics based on such usage remains an open question. However, one possible way forward would be to adapt skipgrams so as to make them able (perhaps with the aid of a large-scale fast surface parser of the kind already applied to large chunks of the WWW) to pick up Agent-Action-Object triples capturing proto-facts in very large numbers. This is a old dream going back at least to Wilks [185] where they were seen as trivial Wittgensteinian "forms of fact," later revived by Greffenstette and Kilgariff [92] as a "massive lexicon" and now available as inventories of surface facts at ISI [86]. These objects will not be very different from standard RDF triples, and might offer a way to deriving massive SW content on the cheap, even simpler than that now offered by ML-based IE. If anything were possible along these lines, then NLP would be able

to provide the base semantics of the SW more effectively than it does now, by making using of some very large portion of the WWW as its corpus. If one finds this notion unattractive, one should demonstrate in its place some other plausible technique for deriving the massive RDF content the SW will require. Can anyone seriously believe that can be done other than by NLP techniques of some type like the one described above?

4.3 The SW and Point-of-View Phenomena

Another aspect of the SW that relates directly to issues of language and language processing is that of a SW as promoting a possible point-of-view on the Internet: so that it would be possible to use filters or annotations to prevent me seeing any web pages incompatible with The Koran, for example, and that might be an Internet-for-me that I could choose to have. However, there is no reason why the SW, or any annotated web should, as some of its critics, such as Ted Nelson have argued necessarily impose a single point of view. He argued that the use of a general ontology necessarily implied seeing only a world compatible with that ontology, and that it might be just plain wrong. The technology of annotations, at least as discussed earlier in this section, is perfectly able to record two quite separate sets of annotation data (as metadata) for the same texts, and no uniformity of point of view is either necessary or desirable. Hendler [76] has stressed the potential of the SW for encapsulating different points of view or even theories in science.

The underlying page-rank technology of Google [137] is itself a point-of-view phenomenon, not in the sense of controlling logical consistency, but as promoting, under a measure of rank, what is most believed, in the sense of being most linked to. This is the basis of the criticism many make of Internet search in general, arguing that what is most believed is not necessarily what is true, quoting as witness the period when, for example, most people are said to have believed the Earth was flat, even though scientists believed it was round, and we may take it as true that it was round at the time. However, it is by no means clear that this is a good perspective on knowledge

in general, particularly in view of an increasing number of phenomena where human aggregates seem better able to predict events than experts, a subject that has been of much interest recently [177] to economists. This has a clear relation to ontologies and other information structures built not from authority but from amalgamations of mass input, sometimes described as the "wiki" movement or "folksonomies" (e.g., http://www.flickr.com/). This movement may, in time, undermine the whole SW concept in a quite different way, since it lacks any concept of "authority" and "trust," but that topic is too broad for discussion here.

The point-of-view and web-for-me issues are of wider importance than speculating about religious or pornographic censorship: they are important because of the notion they imply of there being a correct view of the world, one that the SW and its associated ontologies can control in the sense of controlling the meanings of terms (the subject of this section) as well as the wider issue of consistency with a received view of truth. The nearest thing we have to a received view of truth in the C21, in the Western world at least (and the restriction is important), is that of science, not least because the web was developed by scientists and serves their purposes most clearly, even though they do not now control the Internet as they did at its inception. One could see the SW as an attempt to ensure closer links between the future web and scientific control. This emerges a little in the second view of the SW discussed in the previous section, and its obvious similarity of spirit to Putnam's [148] view that, in the end, scientists are the guardians of meaning.

We do not share this view but we do not dismiss it, partly because meaning and control of information would be safer in the hands of scientists than many other social groups. However, because we believe that meanings cannot be constrained by any such methods, all such proposals, as we argued above, are in the end hopeless.

Let us stay for a moment with a scientific sociology that might be associated with a SW and think what that might mean: on Popper's view [144], for example, scientific activity should be the constant search for data refuting what is currently believed. An SW-for-me for a scientist following a Popperian lifestyle would therefore be a constant

trawl for propositions inconsistent with those held in the scientist's SW. This, one might say, is almost the exact opposite of a normal information gathering style which is to accept attested information (by some criterion, such as page rank or authority) unless contradicted. In the philosophy of science, there is little support for Popper's view on this — a far more conventional view would be that of Kuhn [94] where what he calls "normal science" does not actively seek refutation or contradiction — and we mention it here only to point out that (a) scientific behaviour in information gathering is not necessarily a guide for life and that (b) such differences in approach to novel information can map naturally to different strategies for seeking and ranking incoming information with respect to a point-of-view (or even a personal) SW.

There is another, quite different, implementation of points-of-view that we may expect to see in the SW, but which will be more an application of NLP than a theoretical extension: the idea that many users will need an interface to any web, WWW or SW, that eases access by means of normal conversational interaction, which is itself a classic NLP application area. One can see the need from the Guardian correspondent (8.11.03) who complained that "... the internet is killing their trade because customers seem to prefer an electronic serf with limitless memory and no conversation."

On this view [189] we will need personalised agents, on our side, as it were, emotionally, and over sustained periods: what we have elsewhere called Companions [190]. Their future interest, for the general population at least, is that the Internet is wonderful for academics and scientists, who invented it, but does not serve less socially competent citizens well, and certainly not excluded groups like the old. Some recent evidence in Britain suggests that substantial chunks of the population are turning away from the Internet, having tried it once [47].

As the engines and structures powering the Internet become more complex, it may also become harder for the average citizen to get all they could get from this increasingly complex structure, i.e., the SW, which will be used chiefly by artificial agents. Companions can be thought as persistent agents, associated with specific owners, to interface to the future Internet, and in some ways to protect those owners from its torrent of information about them and for them. These

Companions will know their owner's details and preferences, probably learned from long interaction. Users may also need them to help organise the mass of information they will hold on themselves (see the EPSRC Memories for Life project: http://www.memoriesforlife.org/ or MyLifeBits at Microsoft), and Companions, on the assumption that they will be entities that talk and are talked to, can only be built with NLP technology — specifically dialogue technology.

4.4 Using NLP to Build Ontologies

The previous three sections have all drawn attention to the important possibilities that NLP provides for the development of the SW in a sustainable and tractable way. In this section, we will consider, in some detail, the use of NLP techniques for the development or learning of ontologies, one of the most important instruments in the SW's representation toolbox. Our argument has two parts: we begin by establishing the case for empirically based ontologies, and then demonstrate the practical possibilities by describing a tool for ontology learning.

4.4.1 The Empirical Foundation of Ontologies

We now examine ontologies, as the type of "meaning construct" most central to the concept of the SW, and argue that ontologies are different from thesauri and similar objects, but not in the ways simple definitions suggest: they are distinguished from essentially linguistic objects like thesauri and hierarchies of conceptual relations because they unpack, ultimately, in terms of sets of objects and individuals. However, this is a lonely status, and without much application outside strict scientific and engineering disciplines, and of no direct relevance to NLP. More interesting structures, of NLP relevance, that encode conceptual knowledge, cannot be subjected to the "cleaning up" techniques that the proposers of Ontoclean and other theorists advocate, because the conditions they require can be too strict to be applicable, and because the terms used in such structures retain their language-like features of ambiguity and vagueness, and in a way that cannot be eliminated by reference to sets of objects, as it can be in ontologies in the narrow sense. Wordnet is a structure that remains useful to NLP, and has within it features of both

types (ontologies and conceptual hierarchies) and its function and usefulness will remain, properly, resistant to clean-up techniques, because those rest on a misunderstanding about concepts. The ultimate way out of such disputes can only come from automatic construction and evaluation procedures for conceptual and ontological structures from data, which is to say, corpora.

Is there a problem with ontologies? Are they really distinct from semantic nets and graphs, thesauri, lexicons, taxonomies or are people just confused about any real or imagined differences? Does the word "ontology" have any single, clear, meaning when used by researchers in AI and NLP and, if not, does it matter? And are we, assuming us to be researchers in AI/NLP, just muddled computer people who need therapy, that is, to have our thoughts firmed up, cleaned up, or sorted out by other, more philosophical, logical or linguistic, experts so we can do our job better? And, for those who read newspaper columns on how we are losing the meanings of key words, does it matter that what ontology means now, in our field at least, is somewhere between the notions above, and what it traditionally meant, namely the study of *what there actually is in the world*, such as classes or individuals. For that is what metaphysicians since Aristotle thought it meant.

These questions will be addressed, if not answered, in this section of this paper. The last question concerning whether it matters will, of course, be answered in the negative, since philosophers have no monopoly over meanings, any more than the rest of us. The central question above, as to whether a cleaned up, more logical representation is the way to go, has been a recurrent question in AI itself, and one to which we shall declare a practical, and negative, answer right at the start: namely, that decades of experience shows that for effective, performing, simulations of knowledge-based intelligence, it is rarely the case that enhanced representations, in the sense of those meeting strong, formal, criteria derived from logic, are of use in advancing those ends.

Representational structures are not always necessary where they were deployed in AI, and it is hard to be sure when representations are or are not adequately complex to express some important knowledge, and one should be very wary of the usefulness of logic-based formalisms

where language is concerned. We have already mention the work of Schvaneveldt and Pollack in this regard (cf. Section 2.2.1)

One of the present authors (Wilks [188]) once wrote a paper on MT that was later widely cited, where we argued that most MT systems do not in fact work with the formalisms their architects use to describe them.

The core issues here seem to be: first, whether any particular formalism can encode the requisite types of knowledge for some purpose and do so better than some simpler, more tractable, representation and, secondly, whether that representation supports whatever inference procedures are deemed necessary for deployment of that knowledge. If an ontology is taken in something like its classic sense, as a hierarchical structure of (sets of) entities, properties, relations and events, then we know it will support simple Boolean/quantificational inference and we also know it will have difficulty representing phenomena that do not fall easily under set theory, such as intensional and continuous phenomena. But we also know that these can be accommodated by a range of extensions that make an ontology more like a general AI knowledge representation schema. It seems obvious that, at least in the more formal DAML/OIL parts of the ontology movement, ontologies are formal AI representations renamed with all the representational difficulties identified in the past. Nothing is gained except, perhaps, in terms of the computational tractability of the formalism, a point we shall return to below.

But again, we would stress that appearances are not realities in this area, and much-used ontologies in the public domain do not have the structure they appear to have and would need if the inference constraint were to be taken seriously.

Another theme we wish to introduce is that facts like this last plainly do not make structures useless, because these man-made objects (viz. ontologies, lexicons, thesauri) for classification of words and worlds contain more than they appear to, or their authors are aware of, which is why computational work continues to extract novelty from analysing such objects as Websters 7th, LDOCE, Wordnet or Roget's Thesaurus. Masterman [113] memorably claimed that the structure of Roget showed unconscious, as well as explicit structure, and it was the

mental organisation of a 19th century Anglican clergyman: above all an opposition between good and evil! If any of this is the case, then what structural objects that contain knowledge need is not so much conceptual clearing up but investigation of what they actually contain, and Wordnet has been subjected to a great deal of such analysis, e.g. [140].

As we noted already, those cursed with a memory of metaphysics are often irritated by modern AI/NLP where the word "ontology" is hardly ever used to mean what it used to, namely "the study of what there is, of being in general." Recent exceptions to this are discussions by Hobbs [81] and others, but almost all modern usage refers to hierarchical structures of knowledge, whose authors never discuss what there is, but assume they know all that and just want to write down the relations between the parts/wholes and sets and individuals that undoubtedly exist.

To a large extent, and to avoid pointless controversy, this paper will go along with the usage while noting in passing that as a web search term, "ontology" locates two quite disjoint literatures with few personnel in common: the world of formal ontology specification [84] and the world of ontologies for AI tasks to do with language [109]. Rare overlaps would be Lenat's CYC system, which began life as an attempt to write down a great deal of world knowledge in predicate form but which was also claimed by Lenat as a possible knowledge form for use in language processing. Indeed, the DTO (US) funded the IKRIS program that resulted in IKL, a language that supports interchange between RDF, CYC, etc. Work on video annotation languages (VERL and VEML) has served to draw the AI image understanding community into the SW.

Before proceeding to the critical elements at the end of this section of the paper, it may be worthwhile to set out the basic terms being used, and in particular ontology, dictionary and thesaurus. The following are, in a sense, pure vanilla versions of these:

(1) Figure 4.3 is a bit of ontology for the concept denoted by the English word "pan."

We shall be asking how this is this different from a dictionary entry for the term:

(2) *Pan (N): container, typically made of metal, used for cooking.*

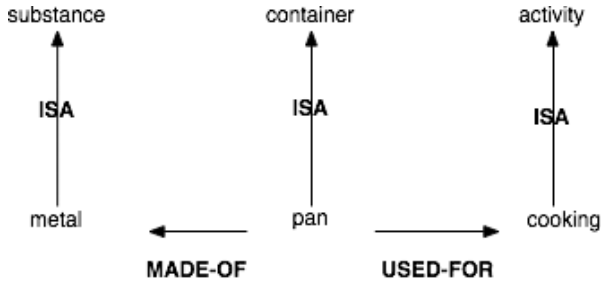

Fig. 4.3 Ontology for the concept 'pan'.

Or for a thesaurus entry for the same word, not in the form of Roget's thesaurus but something more up to date that marks SYN(ONYMS) and HYPER- and HYPONYMS explicitly:
(3) *pan: SYN pot, HYPERNYM container, SYN holder, HYPONYM frying pan, skillet, steamer.*

We shall say that, broadly, (1) is about things in the world, and (2) and (3) are about languages, being, respectively, examples of a conventional dictionary entry and a thesaurus-style lexicon, and the difference will lie in the fact that all the terms in (1), except "pan" itself, are terms in a special defining language.

We could equally well consider:

	GENE	*ALIAS*	*FUNCTION*	*ENZYME*
(4)	*Pb33X*	*PBXalpha*	*activation*	*Transcriptaze Pb33-Xa*

The surface notation in (1) and (4) is different, but this does not matter, since we could easily redraw either in the style of the other. (4) is also about things in the world, genes, not words (other than with the rather special role of names and alias names), and this, in spite of its technical appearance is no more problematic than (1), or is it? We shall come back to genes at the end of the paper and ask whether we can be as sure of genes as we can about pans.

4.4.1.1 Ontologies and Conceptual Relations

The place to begin is one of the questions above about the conflation of ontologies (construed in the modern manner as hierarchical

classification of things or entities) versus thesauri, or taxonomies, as hierarchical classification of words or lexical senses. There is a widespread belief that these are different constructs, as different (on another dimension) as encyclopaedias and dictionaries, and should be shown to be so. Others will admit that they are often mixed in together, as when Wordnet [120] is called an ontology, as parts of it undoubtedly are, on the above definition, but that this may not matter as regards its function as the most popular resource for NLP, any more than it matters that ordinary dictionaries contain many facts about the world like "a chrysanthemum is a flower that grows at Alpine elevations."

A random paper it would be unfair to cite, but which is typical of much recent work, recently offered an ontological coding scheme, made up of what it called Universal Words, and whose first example expression was:

(Drink > liquor)

which was intended to signify, via "universal words" that "*drink is a type of liquor*." This seems at first sight to be the reverse of common sense: where liquors (distilled alcoholic drinks) are a type of drink. Again, if the text as written contains a misprint or misunderstanding of English and "liquid" was intended instead of "liquor" then the natural interpretation of the expression is true. There is probably no straightforward interpretation of the italicised words that can make them true, but the situation is certainly made complex by the fact that "drink" has at least two relevant senses (potable liquid vs. alcoholic drink) and liquor two as well (distillate vs. potable distillate).

We bring up this, otherwise forgettable, example because issues very like this constantly arise in the interpretation of systems that claim to be ontologies, as opposed to systems using lexical concepts or items, and thus claiming to be using symbols that are not words in a language (usually English) but rather idealised, or arbitrary items, which only accidentally look like English words.

This situation, of would-be universal words, taken from English and whose unique sense has not been declared, is sometimes obscured by the

problem of "foreigner English." By that we mean that, since English is now so widely known as a non-maternal language, it is often the case that the alternative senses of English words are simply not known to those using them as primitive terms. This can be observed in the interlinguas in use in some Japanese MT systems. This benign ignorance may aid the functioning of a system of "universal words" but it is obviously an unsatisfactory situation.

We also believe it is not sufficient to say, as someone like Nirenburg consistently does [131], that ontological items simply seem like English words. Our view is firmly that items in ontologies and taxonomies are and remain words in natural languages — the very ones they seem to be, in fact — and that this fact places strong constraints on the degree of formalisation that can ever be achieved by the use of such structures. The word "drink" has many meanings (e.g., the sea) and attempts to restrict it, within structures, and by rules, constraints or the domain used, can only have limited success. Moreover, there is no way out here via non-linguistic symbols or numbers, for the reasons explored long ago in McDermott [118]. Those who continue to maintain that "universal words" are not the English words they look most like, must at least tell us which of the senses of the real word closest to the "universal word" they intend it to bear under formalisation.

One traditional move at this point, faced with the demand above, is to say that science does not require that kind of precision, imposed at all levels of a structure, but rather that "higher level" abstract terms in a theory gain their meaning from the theory as a whole. This is very much the view of the meaning of terms like "positron" adopted by Katz [91].

This argument is ingenious, but suffers from the defect that a hierarchical ontology or lexicon is not like a scientific theory (although both have the same top-bottom, abstract-concrete correlation) because the latter is not a classification of the world but a sequential proof from axiomatic forms.

However, what this "positron" analogy expresses, is very much in the spirit of Quine's [150] later views, namely that not all levels of a theory measure up to the world in the same way, even though there is no firm distinction between high and low levels. This is consistent

with his well-known, but often misunderstood, view that "to be is to be the value of a bound variable," a notion we shall return to later, when we contrast this class of views with those of a writer like Guarino and his co-authors, who very much want to argue that *all* items in an ontology have their own "Identity Conditions," which determine the entities at other levels to which any given node can be linked by a relationship like ISA, normally taken to be set inclusion or type specification.

The most pragmatic possible response to this formalist criticism (like the Ontoclean proposal which we shall discuss in detail below) is to argue that hierarchies, or any other structures, are only justified by their retrieval properties for some desired purpose, and that such evaluation overrides all other considerations. This is a tempting proposal for anyone of a practical bent. However, as a position it has something of the defects of connectionism, expressed by Charniak's remark cited earlier.

Thus even though we, in language engineering, are not doing science in any strong sense, it is still at least aesthetically preferable, of course, to aim for perspicuous, defensible representations, rather than fall back on "it works, shut up about it" as a form of explanation. To the same end, we take the result of many years and rounds of the McDermott discussion (see above) to be that we cannot just say that representations are arbitrary, inessential, and could be unexamined English or even binary numbers. Given all this, one clear way of presenting the issue is to ask the question: are thesauri and taxonomies really different in type from ontologies? Does one describe words and the other describe worlds?

The formal answer is that they are in principle different and should be seen to be so. This is very much the position of the Ontoclean team [53] which we will come to in a moment. However, it is hard to believe they are utterly different in practice, since the principle, whatever it is, is hard to state in clear terms. Facts about words and worlds are often all mixed together.

If Quine [150] is right that analytic and synthetic prepositions cannot be clearly discriminated then it follows straightforwardly that facts about words and the world cannot be separated either. Carnap

[27] proposed a dualism by which a sentence could be viewed in two modes, the material and formal, so as to express both possibilities, roughly as follows:

(F) "Caesar" is a symbol denoting a Roman Emperor.
(M) Caesar was a Roman Emperor.

Carnap's proposal was defective in many ways — the sentences are not synonymous under translation out of English, for example — but was implicated in the origin of Quine's later views, by providing an over-simple opposition, from his teacher, that he sought to overcome and refine.

The position we will defend is the following: the persistent, and ultimately ineradicable, language-likeness of purported ontological terms [131] means that we cannot ever have purely logical representations, purged of all language-like qualities. That enterprise is therefore ultimately doomed to failure, but should be pushed as far as it can be, consistently with the intelligibility of the representations we use [165]. As the history of Wordnet, and its popularity and productive deployment have shown, mixed, unrefined, representations can be useful, a fact formal critics find hard to understand or explain.

It is for this reason that data-mining research has been done for almost 40 years [135] on such human information structures as dictionaries, thesauri, ontologies and wordnets. Were they fully explicit, there would be little for such research to discover.

4.4.1.2 Ontological Basics: A Reminder

Before proceeding to the core of the paper, let us just remind ourselves again of the basic vocabulary for stating the problem: the initial issue in structuring an ontology is to decide what there is out there: are there basically individuals and sets of them or are there also types, concepts, universals, etc., considered as separate entities from individuals, or perhaps as being among them? These are among the oldest intellectual questions known to mankind and there is no settling them. All one can do is make choices and take the consequences.

If one's task is roughly that of Linnaeus — the taxonomy, classification or ontology of the natural species and genera in the world — then

things are relatively simple, ontologically at least. You can assert canaries are birds with a branch structure:

Canary ISA Bird
where the ISA link is set inclusion $\subset$, not set membership $\in$, although one could write
Tweety ISA Canary ISA Bird
as long as the two ISAs are distinguished, thus:
Tweety $\in$ Canary $\subset$ Bird

And this would give us the inference that Tweety is a bird, and Linnaeus could have done this had he been interested in individuals and not only classes. In general, of course, inferences involving $\in$ are not transitive, as $\subset$ is.

Part of the problem with describing an entity like Wordnet is that it does include Linnaean sub-nets taken straight from biology e.g.

Horse $\subset$ Ungulate

If the world consisted only of sets and individuals, then the representational task is over, but the problem really arises when we move to the hierarchical relationships of concepts that may or may not reduce to sets and individuals, and above all, the problems that arise when we mix these, or refer to concepts in ways that do not reduce to claims about individuals and sets of them.

The pathological cases are well known, as in the apparently valid, but actually false, syllogism:

My car is-a ($\in$) Ford.
Ford is-a ($\in$) car company.
Therefore, my car is-a ($\in$) car company.

The problem here can be seen in one of two ways: first, as the illegitimate chaining of ISAs (all of which are $\in$, set membership), and therefore not valid, or, secondly, as the simple word sense ambiguity of the symbol "Ford", which refers to Ford cars as objects in the first line, and to the car company with that name in the second. In the latter case, Ford is some form of conceptual individual, the Ford Car Company and which stands in a relationship to the set of cars it manufactures in a

way that is hard to define, since it is not that of a concept to its corresponding extension set. The conceptual counterpart of a Ford car would presumably be "Ford" meaning something like "whatever it is that makes a car a Ford car". In this case, both premises above are true, but they refer to different objects under the same name, which also leads, by a different route, to an invalid conclusion.

The issues here are subtle and beyond those of the classic one of the duality of intensional concept and its extensional set, diagnosed by Woods in his famous "What's in a link?" paper [199]. One can illustrate the core of the problem for ontologies, which is beyond the Woods duality and reference to fallacies like the one above, by considering a quite different area, treated roughly as follows in Wordnet. One can write down a (partial) taxonomy of religions, say, as follows:

Religion $\rightarrow$ Christianity $\vee$ Islam ...
Islam $\rightarrow$ Sunni $\vee$ Shia $\vee$ Ismaili ...
Christianity $\rightarrow$ Catholicism $\vee$ Protestantism $\vee$ Orthodoxy ...
Protestantism $\rightarrow$ Anglicanism $\vee$ Methodism $\vee$ Presbyteriamism ...
and so on.

This seems something like common sense but it is not at all like a Linnaean taxonomy or an ontology, because it is not clear how to interpret the implied ISA on each arrow link, nor exactly how any node is to be interpreted. If, for example, each node is a set of buildings occupied by a sect, or a set of believers, then the ISA links are set inclusion, provided we can assume the Linnaean disjunction: that no individual or class falls under two or more higher nodes. This is pretty safe in biology (since it is made up that way), though less so in, say, religion in Japan where many are both Buddhists and Shinto believers. But let us assume the disjunction is exclusive so as to continue.

The difficulty is that no such set theoretic interpretation (or model), about buildings or believers, is intended by the common sense interpretation of statements like the ones above, which is usually and loosely described as a taxonomy of concepts. This is sometimes expressed by interpreting the hierarchical links above as PART-OF, as mereology that is to say, which can in turn be seen as either a relationship between concrete objects or conceptual ones or even both, and is (like $\subset$, but

not $\in$, transitive). Thus, one could write (now using $\rightarrow$ to mean "has as part"):

Body $\rightarrow$ Foot $\vee$ Hand
Foot $\rightarrow$ Toe
Hand $\rightarrow$ Finger

to mean "a finger is part of a hand and a hand is part of a body and (so) a finger is part of a body". If we interpret the expressions that way we cannot at the same time be interpreting $\rightarrow$ as $\subset$ since a set of fingers is not a subset of a set of hands. Nor can $\rightarrow$ be interpreted as $\in$ since a toe is not a member of a foot-set. So, the items indicated by the predicates must be "idealised individuals", rather than particular ones, and that is a notion almost identical to that of a concept or intension or sense, namely, what it is to be an individual of a certain kind. Often, what it is that constitutes falling under a concept is the fact that the concept fits below its "mother node" above. Thus "being a finger" is, in large part, being part of a hand, and one could now reinterpret the earlier set of examples in this way, so that "being Catholicism" becomes being part of Christianity. Yet, much of what one wants to put into the definition of a concept does not come from hierarchical relations, of course. So, if we take:

"A Catholic priest is a male."
"A US President is over 35 years old."

These are both true statements about conceptual content that are only incidentally remarks about individuals falling under the description, i.e. the second is true of George W. Bush but it is not a statement about him. The first is currently true, and true till now of the class of Catholic priests, but could change at any time.

The problem of conceptual inclusion, and how to interpret it in a way different from set inclusion or membership relations between objects covered by the concept, is the problem at the heart of the definition of an ontology. Such a structure is different from both a straightforward Linnean taxonomy/ontology (where relations are always set theoretic) on the one hand, and, on the other, from the purely lexical thesaurus like Roget where a concept can be said to fall under another without any analysis or criteria being given for that inclusion.

A move frequently made at this point is to appeal to the notion of possible worlds, and to concepts as picking out, not just a set of individuals in this world, but in some sub-set of all possible worlds. The chief appeal of this move is that it moves entities like the golden mountain to a set of worlds that happens not to include this one, and round squares are said to be found in no world at all. Concepts that would once have been considered as expressing a 'necessary' or 'analytic' relationship, such as "animate cats" then appear in all words, or at least all worlds containing cats. It is often thought essential for formal reasons to constrain possible worlds to a single set of entities, whose (non-essential) properties may change from world to world. This seems an extraordinary constraint on possibility, namely that there is no possible world not containing, say, Tony Blair. Most people would have no difficulty imagining that at all.

This move will not be discussed further here, as it is known to have no computational content, i.e., no process would correspond to searching among all possible worlds. Moreover, if Putnam's [147] arguments have any force at all one cannot know that there are no worlds in which cats are not animate.

In many conceptual ontologies, the concepts are themselves considered individuals, so that set membership and inclusion relations can again be brought to bear, thus yielding the so-called higher-order forms of the predicate calculus. In the religions taxonomy above, we can, quite plausibly, and in tune with common sense, consider Christianity and Islam as individual religions, members of the set Religions. If we then break the former into (sect-like) concepts of Protestantism and Catholicism then, if we wish to retain transitivity, Christianity and Islam will have to become sub-sets of religions and not members of the conceptual class above, at which point the problem returns as to what they are sub-sets of.

4.4.1.3 Is Wordnet an Ontology?

Wordnet is certainly not an ontology in any ordinary sense, and about that its critics are surely right, but it does contain within it a whole range of relations including classically ontological ones such

as set membership and inclusion (what we have called Linnaean). In what follows, we shall concentrate on the Ontoclean formal critique of Wordnet, but the authors of this approach (Guarino, Welty, Gangemi, Oltramari, etc.) are intended only as examples of a school or style, and we could equally well have cited the work of Smith [163, 164].

Wordnet has at least the following relations (not all of which are made explicit):

Linnaean inclusion: ungulates ← horses
Simple subsets: shoes ← tennis shoes
Set membership: painters ← Picasso
Abstract membership: Carcompany ← Ford
Whole-part: Body ← hand ← finger
?Concept–component: Islam ← Sunni
?Concept-subconcept: Physics ← Gravity

As we noted above, Wordnet has many conceptual mother–daughters of the latter kind:

Religion → Islam → Sunni
Religion → Islam → Shia
Religion → Buddhism → Theravada

We also noted already that these cannot be interpreted or modelled by sets of or individual people (though WordNet actually tries this for one sense!), or buildings, etc. It is simply not plausible to interpret any of the above lines as set inclusion relations on, say, religious buildings or people. Because that is not what is meant by anyone who says "Sunni is a major form of Islam."

If therefore one takes the view that an ontology must at its core be a classification modelled by sets, as, say, the Rolls-Royce jet engine ontology is basically reducible to sets of components, or the SmithKlineGlaxo drug ontology to chemical components, then Wordnet is not one, both because it mixes such relations in with others and, most crucially, because its underlying relationship is synonymy, the relationship between members of a synset, and that is certainly not a relationship of sets of objects, and is not an ontological relationship even in the widest sense of that word. It is a logical triviality that one can refer to a concept as well as its instantiations, but this distinction is not

well served if both cannot be ultimately unpacked in terms of sets of individuals in a domain. We noted earlier that the true statement:

A US President is over 35 years old

does not refer to any particular president, but it is easy to state it in a conventional form so that it quantifies over all presidents. But, as we saw, simple quantification does not capture the intended meaning of one who says Sunni is a major branch of Islam, which is a claimed relation of concepts that goes well beyond saying that if anything is a person and a Sunni they are a Muslim.

Wordnet clearly has ontological subtrees, often in the biology domain, yet it cannot be an ontology overall, nor is it a thesaurus, which we may take to mean words clustered by meaning relations, together with major upper level "meaning subsumption" classes e.g. words to do with motion, or games. It is often noted that Wordnet has no way of relating all its elements relevant to the notion of "game", sometimes referred to as the "game problem" in Wordnet (in that tennis, tennis shoe and tennis racquet, for example, are in widely separated parts of WordNet). Yet Wordnet's basic unit, the synset — a list of semi-synonymous words — is very much like that of a row of words in a classic thesaurus like Roget, which also has the "top level" heads and the possibility of cross referencing to link notions such as game.

Interestingly, Roget in his introduction to his thesaurus gave as part of his motivation in constructing it the notion of biological hierarchies, though what he produced was in no way a set theoretic inclusion system. Efforts have been made over the years to provide formal structures that could combine, within a single structure, both set-theoretic inclusions (of an ontology) and the meaning inclusion relations of the kind that typify a thesaurus: the boldest of these was probably the thesaurus-lattice hypothesis of Masterman [113] but that cannot be considered to have been generally accepted.

But the key distinction between Wordnet and an ontology is this: Wordnet has lexical items in different senses (i.e. that multiple appearance in Wordnet in fact defines the notion of different senses) which is the clear mark of a thesaurus. An ontology, by contrast, is normally

associated with the claim that its symbols are not words but interlingual or language-free concept names with unique interpretation within the ontology. However, the position argued here is that, outside the most abstract domains, there is no effective mechanism for ensuring, or even knowing it to be the case, that the terms in an ontology are meaning unique.

This issue is venerable and much wider than the issue of ontologies: it is the issue of the interpretation of terms in all formal systems that appear to be the words of an NL, but where their designers deny that they are. The issue is set out in full in Nirenburg and Wilks [131], a paper in the form of a Greek dialogue where the Wilks character argued that, no matter what formalists say, the predicates in formalisms that *look like* English words, *remain* English words with all the risks of ambiguity and vagueness that that entails. This whole issue cannot be recapitulated here, but it turns on the point of what it is to know that two occurrences of a formal predicate "mean the same" in any representation. For example, it is generally agreed that, in the basic original forms of the LISP programming languages, the symbol "NIL" meant at least false and the empty list, though this ambiguity was not thought fatal by all observers. But it is exactly this possibility that is denied within an ontology (e.g., by Nirenburg and Wilks [131]) though there is no way, beyond referring to human effort and care, of knowing that it is the case or not.

Anecdotal evidence can be sought here in the knowledge representation language CYC, codings in which (2005) have been going on for nearly 30 years and where it is said that there is no way of knowing whether the basic predicates have changed their meanings over that time or not. There is simply no effective method for discussing the issue, as there now is for the ambiguity of word senses in text and their resolution [193].

4.4.1.4 A Programme for Reforming Wordnet as an Ontology

For the last 20 years, Wordnet has been an undoubted success story in providing a resources widely used by the NLP and AI community

[120, 123]. However, it has been severely criticised by academics promoting the Ontoclean approach to ontology engineering [53, 68, 136]. These writers have proposed an alternative approach towards the creation of formal ontologies. We will argue here that this approach is fundamentally misguided and that Wordnet cannot be cleaned up. The approach largely ignores the virtues of Wordnet and the many effective computational uses to which it has been put. Given the argument above, there is no reason to believe that the kind of precision the Ontoclean approach seeks is available for language terms of the kind Wordnet (and, by extension, any ontology) consists of. There is a basic disagreement here about how far NL can be "firmed up" in the way proposed.

The Ontoclean critique of Wordnet begins from observations that it is a mixed bag of representations, a point conceded by its critics and defenders: so, for example, it mixes types for a given initial concept [68]:

apple given as fruit and food (only former is "necessary" according to Guarino [68]):
person given as living thing and causal agent (not "necessarily" the latter according to Guarino [68]).

This is the "multiple appearance" of terms we noted above, and the way in which Wordnet expresses sense ambiguity; it is one of its thesaurus-, as opposed to ontology-like, features. The solution, according to Ontoclean, to the perceived problem is the provision of Identity Criteria (ICs) for concepts: ICs make things what they are and allow their reidentification, which is to say, sufficient conditions for identity, later shifted to necessary conditions, which are supposedly easier to express. On this view, you cannot then hierarchically link concepts with different ICs e.g., ordered sets are NOT sets, a price Ontoclean is prepared, apparently, to pay, since the identity criterion for being an ordered set is quite different from that for being a set. Thus the concept *person* cannot be subsumed by *physical object* because the IC for one is quite different from that for the other. Meeting the sufficient conditions for being a person are sufficient for being a living thing, but not sufficient for being a physical object since *disembodied persons are not impossible*. One sees immediately and from the last case that these issues

will be tricky, even within Ontoclean's frame of reference, and that it may be very difficult to get general agreement, as on the last italicised statement.

But can we have a useful hierarchy if a person cannot be seen to be a physical object? Consider the interpretation of: *Smith tripped and fell on the guinea pig and killed it.*

Bergson [9] in his classic text on humour defined jokes as being occasions when human beings fall under physical laws, as in the standard banana skin scenario, a notion impossible even to state from the Ontoclean perspective since persons do not fall under physical laws, not being physical objects: "... person should not be subsumed by physical object (as a physical object, a body has persistence conditions different from a living being). Yet, these ISA links exist in Wordnet" [136].

Nirenburg's Mikrokosmos ontology [109] has also been a target for reform, and the subsumptions shown in (A) in Figure 4.4 were objected to, and the alternative shown in (B) was proposed, where the lower figure has what are called vertical ontological levels [136]. But one could

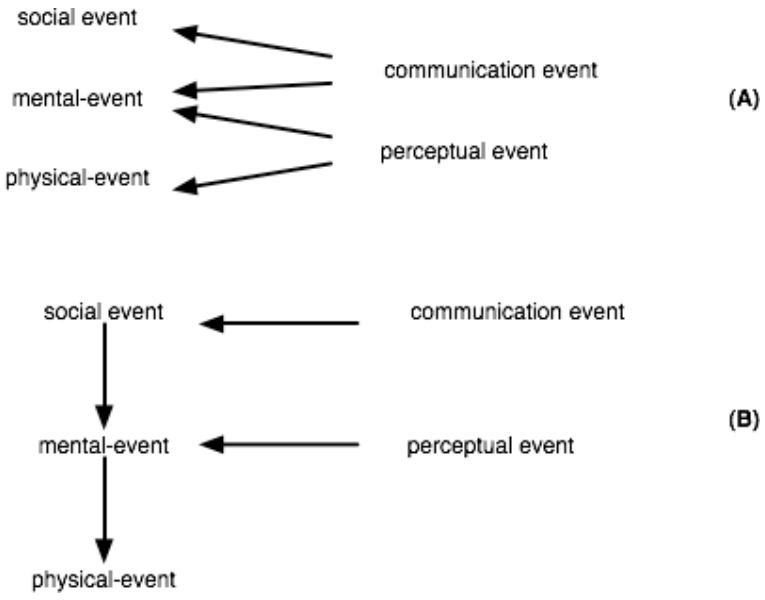

Fig. 4.4 Communication and perceptual events in Mikrokosmos.

still doubt whether any effective ontology could be set up consistently with these principles. Surely, mental events and perceptual events do not have the same identity criteria, an observation made in most elementary philosophy classes, as when one notes that mental events are observable only by one person? If he defines "perceptual event" as subsuming mental events as well as physical, then it cannot be the same kind of entity as Nirenburg was describing in Mikrokosmos in the first place, so it is not clear that the rival mini-ontology covers the same data. Again, if "perceptual event" subsumes both mental and physical events, as appears to be intended by the Ontoclean authors, it is hard to see how its ICs can be the same as, or even compatible with, both sub-events.

One must concede that Ontoclean may well be able to be developed in some less restrictive form of its principles, as has been done in other places, but here we are concerned only with their spirit which may be worse than what we have already shown: if, that is, the issue is really one of differing ICs at each hierarchical level. How can any two hierarchical levels have the same ICs, since, by definition they share features but necessarily have differentia, just in virtue of being different levels? Again, one will expect those differentia to be part of the IC for the appropriate level, in which case how can there be hierarchical relations at all? Canaries have ICs quite different from those of birds, namely *being yellow* among other things. If the IC is taken seriously a Canary cannot be a Bird.

There is of course no intention that Canaries/Birds subsumptions are not in an ontology; our question is how, on current published claims, can this conclusion be avoided in a non-arbitrary way. The problem here cannot be solved just by declaring that ICs need not be the same between levels but only "compatible", since the ICs of items at the same level are also compatible (e.g., wolfhounds and bloodhounds), and one can simply lose any sense of subsumption. There is a also a constant danger of higher nonsense in this area of formalisation; consider: "A piece of coal is an example of a singular whole. A lump of coal will still be a topological whole, but not a singular whole, since the pieces of coal merely touch each other, with no material connection. It will therefore be a plural whole" [136].

It may well be that they really intend "pile" here and this is no more than a slip in their understanding of English, but that is not much cause for comfort because it only serves to confirm our point of how tightly these issues are involved with the understanding and use of the natural language that is more or less isomorphic with the formal language. Yet this point is seldom noticed by the formalisers, it just seems to them an irritant, as opposed to a fundamental issue. On the view advanced here, they take as technical terms, on which to base a theory, words of a language (English) which will not and cannot bear the interpretations required. They cannot express or even admit the senses of the words they use, which is something Wordnet, for all its faults, explicitly allows.

Perhaps this issue is very old indeed: "The Languages which are commonly used throughout the world are much more simple and easy, convenient and philosophical, than Wilkins' scheme for a real character, or indeed any other scheme that has been at any other times imagined or proposed for the purpose". This is Horne Tooke (quoted by Roget at the end of his 1852 Preface to his Thesaurus Roget [155]) attacking Wilkins, perhaps the first ontological formaliser in the 17th century [182].

One way out of this impasse may be something explored by Pustejovsky [146], namely a linking of the ontological levels of Ontoclean to his own theory of regular polysemy. Thus the Ontoclean levels:

mental event
physical event
social event

might be extended to cover classic regular polysemy ranges like:

{Ford = company, car, ?management}

where, for certain concepts, there is a predictable set of functions they can have, and this can be considered as the predictable part of the word-sense ambiguity (alias polysemy) problem. Pustejovsky's approach is not one of multiple entries for a word, in the way a standard lexicon lists a word once for each sense, but a more "compressed" view of a single entry plus lexical rules for its "expansion." However, any link

from phenomena like this to 'ontological levels' would require that the "multifunctional entity" (e.g., Ford) would still appear in more than one place in a structure, and with different interpretations, which would then make it, again, a thesaurus not an ontology.

There is a further issue, much broader than any touched on so far, and which relates to the thrust of much modern philosophy of meaning. There has been general acceptance of the arguments of Quine [150] and Putnam [147], refining as they did earlier positions of Wittgenstein, that it is not possible to continue the two-millennia-old Aristotelean analysis stated in terms of what features a thing or class much have to be what it is: which is to say necessary properties are a dubious notion, and that the associated analytic–synthetic distinction among propositions, deriving from Kant, cannot be maintained.

But there is no hint in the papers on Ontoclean that the authors are aware of any of this, only that there may be practical difficulties — and they concede explicitly that finding conditions, necessary or sufficient, for ICs may be very hard — difficulties in detail, in specifying necessary and sufficient conditions for something's being in a certain class; they never allow that the whole form of analysis is dubious and outdated.

4.4.2 An Approach to Knowledge for the SW

Our approach to ontology learning (OL) has been inspired by a number of different sources. One has been the Quinean view of knowledge which sees it as a space in a state of dynamic tension. Quine believed human knowledge "impinges on experience only along the edges" (Quine [150], p. 39). Statements that came into conflict with experience would result in a re-evaluation or re-adjustment of the statements or logical laws we have constructed, as in any philosophy of science, but he considered that no particular experience is linked to particular statements of belief except indirectly "through considerations of equilibrium" affecting the whole field of knowledge. Thus statements can be held true "in the face of recalcitrant experience by pleading hallucination" (Quine [150], p. 40). This notion of equilibrium and the cumulative (but non-decisive) effect of experience, or in our case textual events, is fundamental to our approach. Each encounter in a text, where one term is juxtaposed

with another, is to be treated as evidence for an ontological relationship. However, it is only the cumulative accretion of sufficient evidence from sources in which the system can have confidence which can be interpreted as ontological facts i.e., knowledge. In this approach there are no absolutes, only degrees of confidence in the knowledge acquired so far.

Another significant impetus has been previous research which has led to a much more subtle understanding of the relationship between a text or collection of texts and the knowledge that they contain. Existing approaches to OL from text tend to limit the scope of their research to methods by which a domain ontology can be built from a corpus of texts. The underlying assumption in most such approaches is that the corpus input to OL is, *a priori*, both representative of the domain in question and sufficient to build the ontology. For example, Wu and Hsu [200] write, regarding their system: "the main restriction [...] is that the quality of the corpus must be very high, namely, the sentences must be accurate and abundant enough to include most of the important relationships to be extracted." In our view, requiring an exhaustive manual selection of the input texts defeats the very purpose of automating the ontology building process. Furthermore, previous research [20] has shown a discrepancy between the number of ontological concepts and the number of explicit ontological relations relating those concepts that can be identified in any domain-specific corpora. This "knowledge gap" problem occurs due to the nature of the texts; they lack so-called "background knowledge," i.e., definitional statements which are explicit enough to allow the automatic extraction of the relevant ontological knowledge. This problem of identifying and capturing essential background information has, of course, been known since the earliest days of the AI "common-sense knowledge" project [100, 116] and is independent of any particular strategy for locating or extracting such knowledge.

In part, a text is an act of knowledge maintenance, not exclusively knowledge creation, in that its intent is based on the assumption the reader will share a considerable amount of common sense knowledge in order to be able to process the text at all. Thus, a given text will at best modify or 'maintain' the knowledge assumed by that text, and, at

worst, may simply repeat it. It is exactly the assumed knowledge which OL wishes to capture, and thus only by having a fuller understanding of the nature of texts can an appropriate methodology be constructed for finding the knowledge available.

4.4.2.1 Requirements

In order to begin to construct a successful OL method, certain observations about knowledge and language need to be taken into account. With due deference to the entirely contrary view espoused by much of philosophy that equates knowledge with truth, the following reflect our view of the problem:

- Knowledge is not monolithic, monotonic or universally agreed. It is uncertain, revisable, contradictory and differs from person to person.
- Knowledge changes continuously over time and so what is believed to be true today will be revised and re-interpreted tomorrow.
- Ontologies are useful if incomplete models of domains.
- Texts assume the reader has a certain amount of background knowledge. The great majority of what we would expect to be in an ontology is exactly in this assumed background knowledge.
- While it is easy to establish that some relationship exists between two terms, explicit defining contexts are relatively rare in texts.

A fundamental premise in our approach is that the set of resources an OL system manipulates — the text, the ontology, and the extraction patterns — are intrinsically incomplete at any given stage, but are, at the same time, the best way for a human to express its view of the domain about which an ontology is to be built, and constitute the best language of interaction with the system. Thus, we believe that the best possible input specification of the task for the OL system to perform is given by a seed ontology, a seed corpus and a seed pattern set. It also follows from the above that we believe that it is not

possible to completely specify the task *a priori* — the ontology engineer should able to (but should not be required to) intervene by pointing out correct/incorrect or relevant/irrelevant ontological concepts and documents, as the process runs, effectively delimiting the domain incrementally through examples. In fact, it is due to the fact that most OL methods expect a task specification as input that they either favour correctness of the acquired knowledge to the detriment of coverage of the domain, or coverage to the detriment of correctness — they are not adaptive with respect to the precision/recall tradeoff. Additionally, given the dynamic nature of knowledge, our approach should allow for the dynamic development of knowledge over time, as more resources are added. Therefore, another fundamental requirement of our approach is for the OL process to be viewed as an incremental rather than an one-off process — the output of one system run can be used as input to another run in order to refine the knowledge. Finally, the data sparsity problem necessitates the use of multiple sources of information.

4.4.2.2 The Abraxas Approach: A Tripartite Model

Our incremental approach is founded on viewing ontology learning as a process involving three resources: the corpus of texts, the set of learning patterns, and the ontology (conceived as a set of RDF triples). Each may be seen in an abstract sense as a set of entities with specific structural relations. The corpus is composed of texts which may (or may not) have a structure or set of characteristics reflecting the relationship between them e.g. they may all come from one organisation, or one subject domain. The learning patterns are conceived as a set of lexico-syntactic textual patterns or more abstractly as a set of functions between textual phenomena and an ontological relationship of greater or lesser specificity. The ontology is also a set of knowledge triples (term–relation–term, or rather domain–predicate–range) whose structure may grow more and more complex as more items of knowledge are collected.

The goal is to create or extend existing resources in terms of one another, with optional and minimal supervision by the user. Any of the three will interact with any of the other two to adapt so that

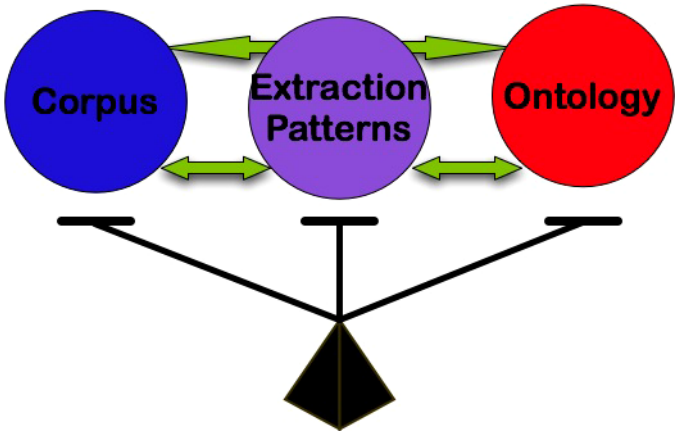

Fig. 4.5 The knowledge aquistion balance: seeking equilibrium.

all three remain in equilibrium (Figure 4.5). The methodology allows, for instance, creating an ontology given an input corpus, extending a corpus given an input ontology or deriving a set of extraction patterns given an input ontology and an input corpus.

The initial input to the process, whether ontology, corpus, patterns or combinations of them, serves both as a specification of the domain of interest and as seed data for a bootstrapping cycle where, at each iteration, a decision is made on which new candidate concept, relation, pattern or document to add to the domain. Such a decision is modelled via three unsupervised classification tasks that capture the interdependence between the resources: one classifies the suitability of a pattern to extract ontological concepts and relations in the documents; another classifies the suitability of ontological concepts and relations to generate patterns from the documents; and another classifies the suitability of a document to give support to patterns and ontological concepts. The notion of "suitability" or rather Resource Confidence (RC) is formalised within a probabilistic framework, one in which the relationship of any resource to the domain is assigned a confidence level (cf. below 4.4.2). Thus, ontology, corpus and patterns grow from the maximum probability core (the initial or manual input) to the lower probability fringes as the process iterates. This is conceived

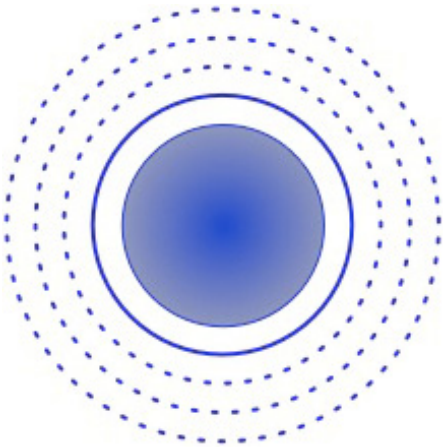

Fig. 4.6 The central core resources surrounded by layers of diminishing resource confidence.

as a set of concentric circles of diminishing RC (cf. Figure 4.6), where the position of any item is continuously changeable depending on the addition of newer items. Newly added elements have an initial neutral probability but as the data they provide is further confirmed confidence in them increases (or decreases) accordingly. This means that subsequently added elements over time can acquire a confidence level equivalent to the manual input. Stopping criteria are established by setting a threshold on the lowest acceptable RC for each resource type, or by setting a threshold on the maximum number of iterations, without any new candidate resources for each resource type being obtained. The incompleteness of the corpus is tackled by iterative augmentation using the web or any other institutional repository as a corpus. Corpus augmentation in our approach consists of a set of methods that aim to incrementally add new documents to the corpus, such that documents with higher relevance to the domain are added first. We present further details about certain aspects of this model in the remaining sections below.

4.4.2.3 The Corpus

As we have argued elsewhere, the corpus is the fundamental source of knowledge in text-based ontology learning. It provides an entry into

the mind of the authors of texts, and by collecting the documents of a institutional community or subject domain, provides significant insight into the common vocabulary and background knowledge of those communities or domains.

We conceive of the corpus as an open-ended set of documents. This document set may be empty initially, partially filled or filled with a considerable set of documents. In each case, this has different implications for our approach:

- **Empty set.** In such a case, the domain of the ontology must be specified by a seed ontology of greater or lesser size. The greater the seed ontology the more rigourously defined the domain. Documents will be added to this empty set in response to retrieval queries generated from the ontology terms.
- **Partially populated set — the seed corpus.** In this case, the documents are seen as defining the domain, either alone or in tandem with the ontology. The domain is defined by the terms or key words in the corpus.
- **Full corpus.** In principle, the ABRAXAS model assumes that all corpora are open-ended but in some cases the model needs to work with a corpus that is defined as all there is i.e. no further documents are available. This will have an impact on the degree of detail that an ontology derived from the corpus will have.

In the first two the expansion and growth of the corpus is achieved by iteratively adding further documents which fulfil certain criteria, typically of providing explicit ontological information (as defined in [20]). The seed corpus can have a further function in playing a role in the stopping criterion. Thus if the ontology grows to the stage that it covers the Seed Corpus adequately, then this can be identified and halt the process (cf. Section 4.4.2).

Documents can be added to the corpus from a variety of sources. The web is an obvious source and has been widely used in ontology learning work. An institutional or organisational repository may be another source. Furthermore, there may be specific external sources

which may be defined to be used (for example, for reasons of provenance or trust). In such a case, only texts of a certain genre or chronology would be made available to the system. A key design issue is how documents are selected and queued to enter the corpus. Should a new document be needed, a query is constructed for any one of the possible sources of new documents. The most typical situation in which an additional document needs to be added is when there is an absence of explicit knowledge defining the ontological relationship between two terms. A document query is constructed containing the relevant extraction pattern and query terms, and documents which contain these are identified. These are queued for possible addition to the corpus.

Each document has an associated Resource Confidence level (RC), range [0,1] (explained further in Section 4.4.2.9). A document existing in the seed or full corpus will have an RC usually of [1.0], although the user may choose to set another value. A document added subsequently will begin by having a confidence level just above neutral [0.5] because it will have fulfilled some criteria for inclusion. If it proves to have a set of terms very close to the seed corpus, confidence increases and equally if a number of extraction patterns are applicable again confidence in this document increases.

4.4.2.4 The Ontology

As noted above, the ontology, much like the corpus, consists of a set of objects, in this case knowledge triples (term, relation, term or in RDF terminology domain, predicate, range). It is also an open-ended set which initially may be more or less fully specified. Thus like the corpus, the ontology set may be empty, partially filled or fully filled set of triples.

- **Empty set.** The ontology in this case provides no guidance to the ontology building process. Its size and scope will be entirely determined by the other two entities in the system i.e. the corpus and the extraction patterns.
- **Partially filled set.** This situation arises typically where the ontology is acting as a 'Seed Ontology' which will

determine the domain for the ontology building process, or where an ontology exists already and the system is being used to update or expand the ontology.

- **Fully filled set.** This situation arises only if the user believes the ontology fully represents the domain and wishes either to expand the document set (using the ontology as a complex source of queries over the domain) or as a bootstrapping knowledge source for the extension of the extraction pattern set.

In the first two cases, knowledge triples are added iteratively as they are obtained by applying the extraction patterns to the corpus. A seed ontology can act as input to the stopping criteria if for example the system is required to identify textual evidence for confirmation of the existing knowledge in the ontology and asked to stop once that has been achieved i.e. once the corpus covers the ontology. The partial and fully filled ontology sets can be used to learn additional extraction patterns from the corpus (existing or obtained).

Knowledge triples are added to the ontology set as a result of either manual insertion or as an output of the application of the extraction patterns to the corpus of documents. Manually inserted triples have a confidence level of [1.0] while automatically added triples have a confidence level which is a function of the document from which they are derived and the reliability of the extraction pattern. Over time as additional documents are added to the corpus, the confidence levels for any given triple may increase and be adjusted accordingly.

There are three important points in which our conception of the ontology set differs from an ontology as usually envisaged. First, a set of triples does not add up necessarily to an ontology as usually conceived. We envisage a post-processing stage which rationalises the multiple triples and constructs a hierarchy from the multiplicity of triples. This in itself is not a trivial task but it also reflects the cognitive reality that we may have disparate items of information, individual facts, which do not necessarily add up to a coherent whole. Only a sub-part of the ontology set may actually be rationalisable into a normalised ontology. Secondly, confidence levels are associated with each triple and these

change continuously as further data is processed by the system. These confidence level start a neutral level for every possible term–relation–term triple (0.5) and alter in the light of further evidence both positive and negative. As more and more evidence for a knowledge triple is encountered so the confidence level increases. As evidence mounts that any two terms are never encountered together, so the confidence level decreases for a given candidate triple. Thirdly, we also envisage an neutral ontological relation so as to be able to state the fact that terms T_x and T_y and closely related by some R but the type of that relation is unknown. This allows the ontology to reflect relational information based on term distribution (of the type identified by term association algorithms and sometimes represented by PathFinder Networks) while still capturing the absence of knowledge concerning the specific nature of the ontological relationship.

4.4.2.5 The Extraction Patterns

Lexico-syntatic patterns are specific detailed templates which match a sequence of words and annotations — typically POS speech tags but not necessarily. They have an extensive history in NLP having been used for a variety of purposes including QA [157] and obtaining facts from the web [48]. However, we would like to propose that it is merely the most precise from a range of *Extraction Patterns* all of which obtain certain information from the text concerning the relationship between terms in the text, or between terms and something else (such as the corpus or an ontology). Under this more abstract view, we see a hierarchy of extraction patterns (EP), starting from a very simple level and moving to a more complex and more ontologically precise level.

The most basic level (or Level 0) is an EP which identifies that a given term is specifically related to the domain. We use the term *domain* loosely and depending on the purpose a level one EP identifies a term as belonging to the corpus, to the subject area, or to the ontology. From this perspective, term recognition is the basic level one EP. All that term recognition does is establish that there is greater relationship between a certain set of terms and the domain (ontology, etc.) than with the language as a whole. Typical term recognition methodologies

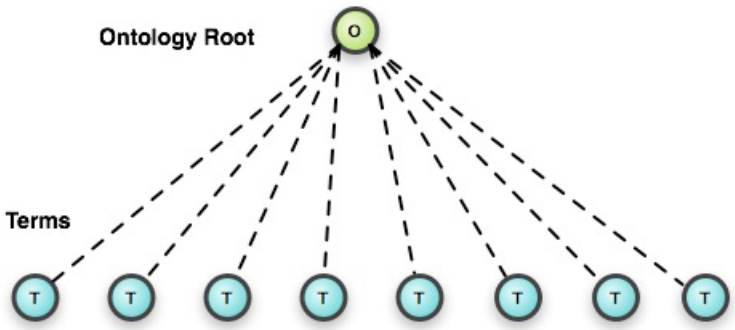

Fig. 4.7 The result of applying the Level 0 extraction pattern (term recognition).

use a reference corpus to identify the most discriminating terms in the domain corpus [2, 178]. Of course, a variety of methodologies can be applied to perform term recognition. The output of such an EP can be visualised as shown in Figure 4.7.

At the next level in the hierarchy, Level 1, EP establish that there is some sort of relationship between two terms and that relationship is greater than with other terms. Of course, such term association algorithms will often rank the degree of association (cf. more detailed discussion in [17]). The establishment of some sort of relationship is important both in limiting the work of an ontology engineer (i.e., guiding their input) and more importantly in the context of applying EPs at Level 2 (lexico-syntactic patterns) in limiting the search space. If a corpus has (only) 100 candidate terms, then theoretically there are 9900 bigrams to potentially consider for possible relations. This set of pairs with potential ontological relations can become rapidly unmanageable if we cannot select the best candidates. Thus, following the application of level 1 EPs, we might theoretically obtain a situation as shown in Figure 4.8 (ignoring issues of data sparsity and a number of other issues, for the moment).

Finally, at the last level (level 2), lexico-syntactic EPs specific to particular ontological relations can be applied. These EPs are of the type discussed above and represent the most explicit recognition and definition of ontological relations. Thus for each ontological relation (ISA, PART_OF, LOCATED_IN, etc.) a set of EPs apply, and where

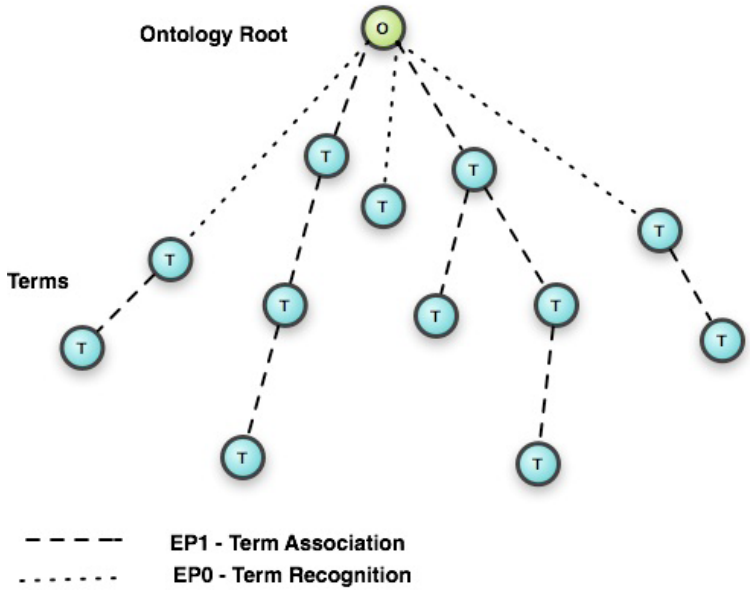

Fig. 4.8 The result of applying the Level 1 extraction pattern (term association).

they do apply successfully then the ontological relationship between the terms can be labelled. This will tend to be a subset of the terms identified and associated in the earlier steps, resulting in something like Figure 4.9.

At each level, some degree of knowledge is being extracted from the texts. For example, the level 1 structures of purely word association capture significant knowledge [161] even if the lack of labelling still means that the precise nature of the knowledge is still vague. Following the application of relation labelling, we clearly reach a degree of explicitness where reasoning systems can be used more effectively.

As in the other two entities in the Abraxas approach, the EP set is an open-ended set. Clearly, it is more limited in size and scope than the other two sets and this is due to the relatively limited number of explicit lexico-syntactic patterns which exist in NL. Nonetheless, new patterns do exist and develop and there are frequent domain specific ones either related to a specific subject area (e.g. legal documents) or to specific types of texts (glossaries or dictionaries for example).

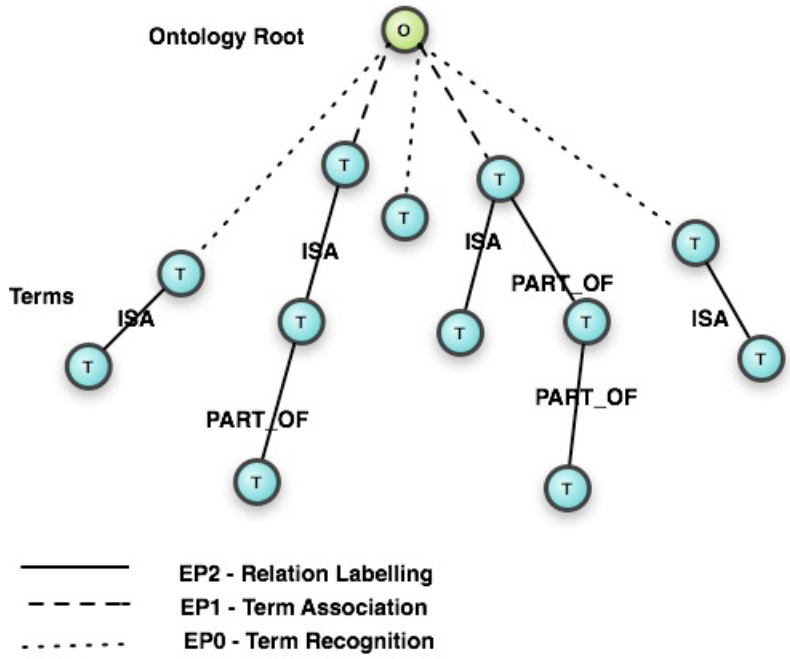

Fig. 4.9 The result of applying the Level 2 extraction patterns (ontological relation recognition).

The extraction pattern set, like the other two entities, may be empty, partially filled or fully filled:

- **Empty set.** This situation is essentially counter-intuitive in that it means that the system has no predefined means for recognising ontological relations and is entirely dependent on the information provided by the corpus and the ontology. We believe that in principle all extraction patterns should be learnable by the system from the combination of ontology (which provides examples of terms related by labelled relations) and the corpus (which provides examples of instantiations of the knowledge triples in texts).
- **Partially filled set.** This is the typical starting situation. As described above, we assume a default extraction pattern essentially identifying the existence of an ontological

association between two terms, and a number of sets of lexico-syntactic extraction patterns. Each set corresponds to a specific ontological relation. Furthermore, each set may expand by identifying further instantiations in text of two terms.

- **Fully specified set.** In this case, all extraction patterns the system is allowed to use are predefined. This may arise if the user wishes to obtain very high precision or only extract a specific type of ontological relation.

It should be clear that we consider the first and third case described here as being exceptional. The standard situation involves a predefined initial set of extraction patterns which are expected to grow over time with exposure to further data. The set of extraction patterns act as stopping criteria because if nothing new can be extracted from the corpus then either (a) a new extraction pattern must be added or (b) a new document must be added. If neither of these is justified (e.g., due to the absence of explicit knowledge) then the system can halt.

New extraction patterns are added when a certain ontological relation is instantiated repeatedly in a specific lexico-syntactic environment. This presupposes that the ontological relation has already been made available to the system and example knowledge triples exist. The following steps then occur: (a) Given a knowledge triple of the form $\{T_x\, R\, T_y\}$ a number of sentential contexts (citations in corpus linguistic terminology) are identified; (b) the citations are analysed and abstracted over in a manner similar to that used in the adaptive IE algorithm LP^2 [34]; and (c) the abstracted lexico-syntactic pattern is added to the EP set with a given confidence value dependent on how many existing (known) knowledge triples it actually extracts.

4.4.2.6 The Fourth Dimension: The System User

There is a fourth aspect to these three resource sets and that is the system user. Of necessity, the system user must provide some initial seed resources, whether in the form of a corpus, a set of extraction patterns, or a seed ontology. As noted above, there is no need for the user to provide more than just some basic seed data because the system is

designed to re-assert its equilibrium by seeking compensatory resources from elsewhere. The consequence of this is that the user must also specify what external resources the system has access to. These may include the whole web, some particular subpart (for example, a range of IP addresses), a specifically defined repository, a corporate intranet, etc.

When the user defines manually some specific resources, they are designated as having a specific degree of confidence (RC). There are two options, as the model currently stands. In one case, the user specifies the resource as maximally belonging to the domain (RC = [1.0]). In the other case, the resource is defined as not belonging to the domain (RC = [0.0]). This allows the user to provide both positive and negative inputs for the start of the process, and also in the course of the running of the system it will be possible to designate certain resources as [1.0] or [0.0] and thereby 'guide' the system as it iteratively progresses in its process of knowledge acquisition.

4.4.2.7 The Basic Method

The basic or core method in Abraxas involves an iteration over the three resources mentioned above triggered by some measure of in-equilibrium. The starting state necessarily involves some initial manual input to the resources. The system will then test for "equilibrium" as mentioned above, i.e. whether the corpus provides evidence for the knowledge triples given the EPs, whether the extraction patterns provide evidence for the knowledge triples given the corpus, whether the knowledge triples provide evidence for the corpus given the EPs. At an abstract level, we can visualise this process as shown in Figure 4.10. A full description of the system can be found in [17].

4.4.2.8 Metrics for Measuring In-Equilibrium

Our approach to managing knowledge is based on using a number of metrics to measure the difference between resources and the confidence levels attached to each resource. The difference measures concern the difference between what the resource provides and its fit with the other resources. Thus, a corpus may have more terms or keywords in it than present in the ontology — this defines a system level in-equilibrium

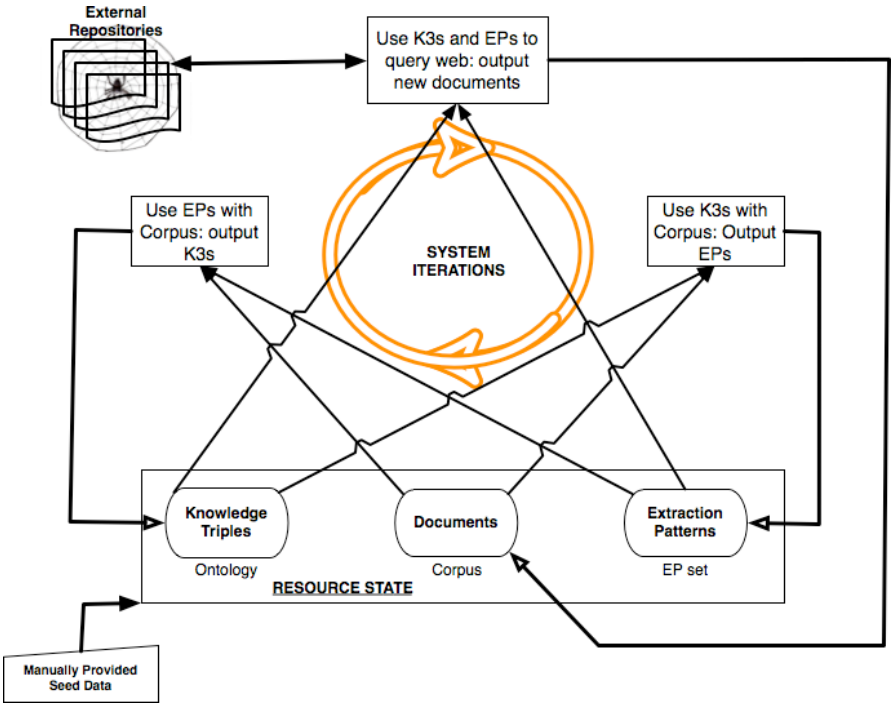

Fig. 4.10 An abstract view of the Abraxas system.

and thus triggers certain actions. The confidence level measures also provide a measure of the system state and thus may trigger a certain action.

These metrics are used together with external measures, such as defining the number of iterations, to determine when the system will either stop or request input from the user. We describe the Knowledge Gap and explicit Knowledge Gap measures below. In the longer term, we believe a number of metrics need to be identified which reflect the internal dynamics of knowledge processing. For example, how do human beings decide when they have enough knowledge concerning a given subject? How many examples of the usage of a term are needed for comprehension of the term and consequent addition to the body of knowledge? How much contradictory data is needed before a

piece of knowledge is rejected entirely? These are questions for further research.

4.4.2.9 The Resource Confidence Measure

A key measure in the Abraxas system is the resource Confidence Measure (RCM) which measures the confidence that the system has in a given resource (e.g., document, extraction pattern, or knowledge triple). From an intuitive perspective, the RCM provides a means for the confidence in any given item of knowledge to be determined, and in a complementary fashion the confidence for any given extraction pattern or document. The RCM value for a knowledge triple reflects how confident the system is that it is a correct piece of knowledge (in the sense that the system has found evidence for its assertion in the text). The RCM value for an extraction pattern reflects how confident the system is that the extraction pattern will extract accurate pieces of knowledge. The RCM value for a document reflects how confident the system is that the document belongs to the domain of interest and provides valid knowledge triples by providing at least one sentence where an extraction pattern matches.

As noted above, in the Abraxas approach, manual intervention can take the form of specifying that a resource belongs to the domain (and therefore has an RCM [1.0]) or that it does not belong (and therefore has an RCM [0.0]). Equally, the user could specify that they do not know or have an opinion about a given resource and specify it as having an RCM of [0.5]. However, system added resources, whether documents, knowledge triples or extraction patterns are assumed to have varying degrees of confidence associated with them. The confidence value is a function of the success or suitability of a given resource in deriving the corresponding other resource. For example, we evaluate a given EP in terms of known knowledge triples and documents. If the precision of an extraction pattern is high in the known/learning context, then we can assign it a high confidence value so that the system can be confident that any further knowledge triples extracted by that EP will be correct.

In a similar manner, we can assign a certain RCM to a knowledge triple depending on its source. If it has been added to the ontology set manually its RCM is [1.0], if it has been identified by using a certain EP over a specific set of texts, then the confidence value is a function of the confidence values assigned to the EPs and documents involved. Equally, for a given document, we can assign a confidence value dependent on how fitting to the ontology the knowledge triples extracted are given the existing EPs. Thus, for each resource set, confidence for any resource item is defined in terms of the other resource sets. This means that for any given resource, there is a corresponding set of resource pairs with which it interacts.

4.4.2.10 The Knowledge Gap Measure

Let us suppose we have defined initially a seed corpus, which may be either a given set of domain texts or a set of texts retrieved using a seed ontology to provide the query terms. Whether retrieved from the Web or from a specific text collection is immaterial. At an abstract level, this seed corpus contains a certain quantum of knowledge, or more accurately, in the light of Brewster et al. [20], one must assume background knowledge which is the relevant ontology we wish to build. The *knowledge gap* (KG) is the difference between an existing ontology and a given corpus of texts.[2] The knowledge gap is measured by identifying the key terms in the corpus and comparing these with the concept labels or terms in the ontology. Such an approach depends on being able to successfully identify all the key terms in a given corpus and also assumes that all the key terms should be included in the ontology.[3]

We define the KG measure as the normalisation of the cardinality of the difference in the set of terms in the ontology (Ω) vs. the set of terms in the corpus (Σ) as shown below:

$$\mathrm{KG} = \frac{|(\Omega \cup \Sigma) \backslash (\Omega \cap \Sigma)|}{|\Omega \cup \Sigma|} \tag{4.1}$$

[2] This is closely related to the notion of 'fit' we have proposed elsewhere [18].

[3] In practice, we believe this to be a relative scale i.e. terms are recognised as key terms and therefore to be included in the ontology with varying degrees of confidence.

At the beginning of the ontology learning process, KG will initially be either 1 (maximal) or very close to 1, indicating a large 'gap' between the knowledge present in the ontology and that which needs to be represented which is latent in the corpus. As the ontology learning process progresses, the objective is to minimise KG as much as possible while realising, in accordance with Zipf's law, that this is an asymptote.

There are a wide variety of methods for identifying the salient or key terms in a corpus of texts [1, 115] but the real challenge is to learn automatically the ontological relationship between terms [22]. It is also relatively uncontentious to use distributional methods to show that there *exists* a relationship between two terms, and in a large proportion of these cases that relationship will be an ontologically relevant one. We define *explicit knowledge* as textual environments where ontological knowledge is expressed in a lexico-syntactic pattern of the type identified by [75]. In such a case, the ontological relationship (domain–predicate–range) is automatically extractable. However, for any given corpus, a minority of the terms will occur as explicit knowledge and thus can be automatically added to the set of ontological knowledge. The difference between the set of pairs of terms whose ontological relationships are known (because they are in the ontology) and those which need to be added to the ontology (because they are key terms in the corpus) and whose ontological relationship is unknown, we will term the *Explicit Knowledge Gap* (EKG). The absence of explicit knowledge may be between two terms both absent from the ontology or between a new term and a term already assigned to the ontology set. Let Π be the set of triples based on terms existing in the corpus, which are known to have some kind of ontological relationship on distributional grounds. Each triple consists of t_i, r, t_k where $t_i, t_k \, \epsilon \, T$ the set of all terms in the corpus, and $r \, \epsilon \, R$ the set of ontological relations. The set Π consists of triples where the ontological relationship is unknown (let us represent this by r_0). In contrast, Ω_3 is the set of pairs of terms whose ontological relationship is explicit, and these are assumed to be the set of knowledge triples in the ontology. Again obviously these can be represented as triples consisting of t_i, r_j, t_k where, in this case, $t \, \epsilon \, \Omega$ i.e. the set of terms in the ontology.

Thus EKG is defined analogously to Equation (4.1), where Ω is replaced by Ω_3, and Σ is replaced by Π:

$$\text{EKG} = \frac{|(\Omega_3 \cup \Pi) \backslash (\Omega_3 \cap \Pi)|}{|\Omega_3 \cup \Pi|} \tag{4.2}$$

While EKG will never be 0, in a similar manner to KG, one objective of the system is to minimise this EKG. The seed corpus will have relatively high KG and EKG measures. The expansion of the corpus occurs in order to reduce the two respective knowledge gaps and consequently learn the ontology. As this is an iterative process we can conceive of the expanding corpus and the consequently expanding knowledge like ripples in a pool (Figure 4.6).

The exact formulation of these equations will change and further metrics are likely to be added as our understanding develops of what it means to represent uncertain knowledge and the degrees of uncertainty. For example, we are still unclear as to how to model correctly the confidence values which depend on the nature of the source of the knowledge. How do we globally make knowledge derived from source X more confidence worthy than source Y, especially if the latter has greater quantity so a given knowledge triple may be derived more frequently and yet may be incorrect. This raises the issue of trust of internet based sources, a topic researchers are only beginning to understand [3, 134].

4.4.2.11 Some Evaluation Results

Ontology evaluation is a challenging topic in itself because knowledge cannot easily be enumerated, catalogued or defined *a priori* so as to allow for some sort of comparison to be made with the output of ontology tools. Various proposals have been made in the literature [59, 181]. An evaluation by Gold Standard (GS) was chosen here. For that purpose, we created a domain-specific hand-crafted ontology reflecting the basic common sense knowledge about animals which a teenager might be expected to possess, containing 186 concepts up to 3 relations deep.[4] In order to compare the GS ontology with the computer generated one,

[4] Publicly available from http://nlp.shef.ac.uk/abraxas/.

we chose to follow the methodology proposed by Dellschaft and Staab [45]. The following metrics are thus used: Lexical Precision (LP) and Recall (LR) measure the coverage of terms in the GS by the output ontology; Taxonomic Precision (TP) and Recall (TR) aims to identify a set of characteristic features for each term in the output ontology and compare them with those of the corresponding term in the GS; Taxonomic F-measure (TF) is the harmonic mean of TP and TR, while Overall F-measure (TF′) is the harmonic mean of TP, TR, LP and LR.

A series of experiments were conducted, each time varying the seed knowledge input to the Abraxas system. In all cases we used as a stopping criterion the EKG measure described above. We used the same set of six extraction patterns, shown in Table 4.1, which previous research had shown to have good precision. Pattern learning was disabled in order to separate concerns — we intended to isolate the ontology learning process from the influence of pattern learning in these experiments, making results more comparable with those of the literature. For the same reasons, the system was tested in a completely unsupervised manner.

The parameters varied were (a) whether Wikipedia texts were included as part of the seed and (b) the number of seed RDF triples. In the cases where we used a set of texts as part of the seed, we chose an arbitrary set of 40 texts from Wikipedia all of which were entries for commonly known animals such as *hedgehog, lion, kangaroo, ostrich, lizard*, amounting to a little over 8000 words. When more than one triple was included in the seed set then it was a random selection from the GS. The following list shows the different cases:

Experiment 1: Corpus: 40 Wikipedia texts; Ontology: {dog ISA animal};
Experiment 2: Corpus: 0 texts; Ontology: {dog ISA animal};

Table 4.1 Extraction patterns used: NP = noun phrase, sg = singular, pl = plural.

NP(pl) such as NP*	NP(sg) is a kind of NP(sg)
NP(sg) or other NP(pl)	NP(sg) is a type of NP(sg)
NP(pl) and other NP(pl)	NP(pl) or other NP(pl)

Experiment 3: Corpus: 0 texts; Ontology: {worm ISA animal, bird ISA animal, reptile ISA animal, rat ISA mammal, dog ISA mammal};

Experiment 4: Corpus: 40 Wikipedia texts; Ontology: {worm ISA animal, bird ISA animal, reptile ISA animal, rat ISA mammal, dog ISA mammal};

4.4.2.12 Comparison with the Gold Standard

Our initial experiment was with Case 1, running over approximately 500 iterations. The final results are shown in Table 4.2. Both the TF and TF′ obtained are comparable with equivalent results in the literature, which often achieve maximum scores around [0.3] for both precision and recall.[5]

4.4.2.13 Learning Curves

Figure 4.11 shows how the results vary over the number of iterations. We can see here that LR steadily increases reflecting the growing size of the ontology and correspondingly its overlap with the GS. In contrast, LP is in constant flux but with a tendency to decrease. TP varies between set limits of [1.0–0.84] indicating that concepts are generally inserted correctly into the hierarchy. TR is also a measure in considerable flux and manual analysis of the different output ontologies show that sudden insertion of parent nodes (e.g., mammal at iteration 9) make a substantial difference which gradually stabilises over further iterations. Over long numbers of iterations, this flux in TR seems to become less likely. We also observe a steady increase TF′ in parallel with the increase in LR indicating that the system is doing better as it increases its coverage of the lexical layer of the GS ontology.

Table 4.2 Results obtained for experiment 1.

LP	0.40	LR	0.48
TP	0.95	TR	0.70
TF	0.81	TF'	0.60

[5] For example Cimiano et al. [33], but note the ontology and the details of the evaluation methodology were different.

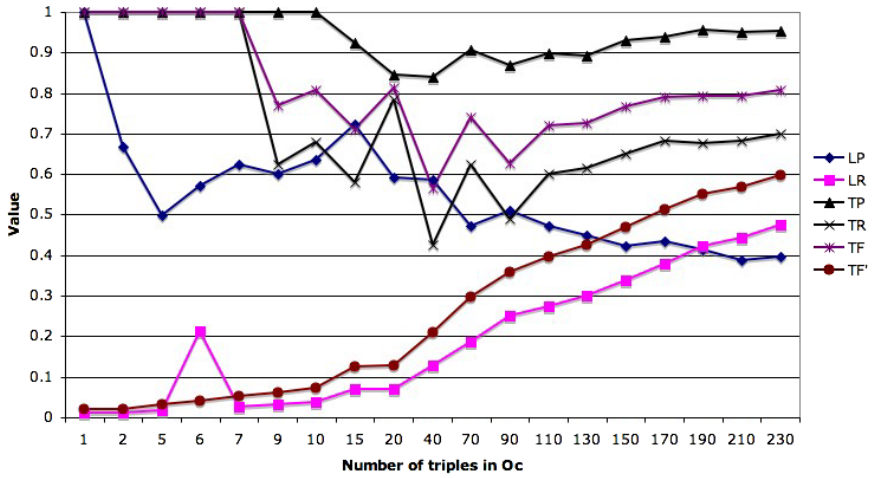

Fig. 4.11 Evaluation measures (LP, LR, TP, etc.) plotted against the sequentially produced ontologies from the iterative process.

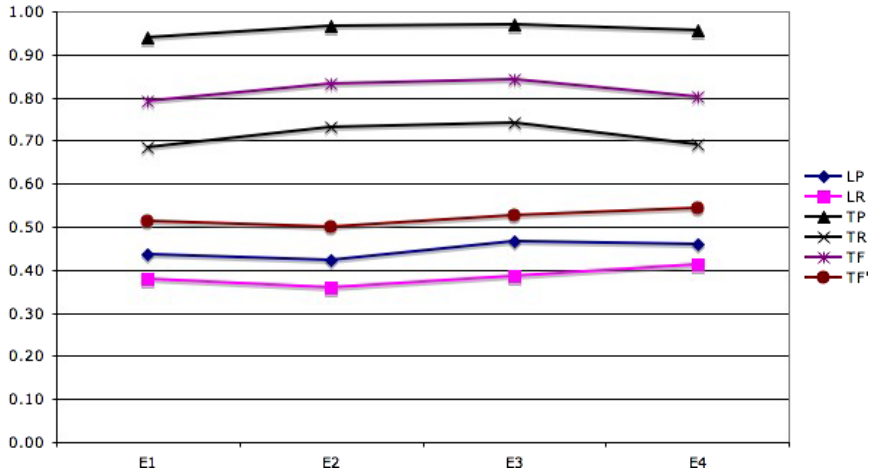

Fig. 4.12 Comparison of final evaluation measures for the different seed cases (experiments 1–4).

4.4.2.14 Variation of Initial Seeds

In order to evaluate the effect of varying the seed data, we then ran the same experiments with the three other cases listed above. The results are shown in Figure 4.12. By comparing cases 2 with 3, we see that the

number of triples used to start which has little impact over the final results; by comparing cases 1 with 2, and cases 3 with 4, we see that TR drops slightly when the starting seed contains texts.

The low LP and LR do not accurately reflect the real quality of the generated ontology because a manual inspection of the ontology showed that in 230 triples, there were 225 concepts of which only 14 could be clearly seen to belong to another domain (flour, book, farmer, plant, etc.), and another 10 were borderline (predatory bird, domestic dog, wild rabbit, large mammal, small mammal, wild fowl, etc.). So a manual evaluation would suggest 201 correct terms or [0.89] precision. LP has a tendency to decrease, which is to be expected as the system is using the Web as a corpus so it will inevitably include items absent from the GS. The difference between the different seed cases is not very great. The system tends to produce consistent results.

Our approach could be criticised for using an ontology which is too simple, but our intention is to capture common sense knowledge which tends not to include abstract concepts. Initial experiments with more complex ontologies indicate a degradation in relative performance. Our current system uses a very simple approach to term recognition and using a more sophisticated approach will increase both our recall and overall performance. At this stage in this work, the knowledge acquired is largely taxonomic knowledge, the knowledge acquired is exclusively taxonomic, but the general principles should be applied in future to a wider range of types of knowledge.

5

Conclusion

The Abraxas model of ontology learning does not depend for its validity on the details of its implementation. As we have noted above, we view Abraxas as a meta-model whose key features include the treatment of resource confidence as uncertain and the corresponding ability to model uncertain knowledge, the use of multiple resources for knowledge acquisition, the ability to seed the system with varying degrees of complex data (texts, extraction patterns and knowledge triples/fragments of ontologies) but above all the notion that the system strives for an equilibrium between the three resources.

Ontology learning makes it possible to capture a certain amount of knowledge that resides in texts. But an effective approach which can utilise this kind of technology needs a re-appraisal of how knowledge is represented and used. The current model which dominates the SW is one where knowledge is represented in a subset of FOPL called Description Logic as expressed in the OWL language. As long as knowledge is represented in this formalism "reasoners" such as FACT and Pellet [85, 162] can be used to perform various logical tasks such as determining set inclusion, inconsistency and in some cases forms of knowledge

discovery [106]. There are severe limitations to functioning solely in the world as delimited by OWL and the corresponding reasoners [172].

There appears to be a coherent case to be made for research efforts to move the NLP automatically generated knowledge representations, on the one hand, and the formal hand crafted ontologies, on the other, towards a middle ground where automatically generated ontologies succeed in being more fully specified while concurrently the inferential and reasoning models we use allow for "dirty" knowledge representations which are approximate, uncertain and more typical of human language in its power and flexibility. The reliance on very formal and logically specified ontologies is one reason why Berners-Lee favours the exclusive use of curated data. There is also a strong case to be made to treat "folksonomies" and "tag clouds" as further examples of dirty knowledge representations which need to be used to the full (a number of authors are currently taking up this idea Gruber [67], Damme et al. [43]). In reality, we need to move to a more sophisticated approach towards knowledge representation. This is absolutely necessary in the world of the SW and is just as needed in more specialised domains such as systems biology where knowledge changes continuously.

We can conceive of presenting knowledge somewhat as shown in Figure 5.1 where the body of knowledge is conceived of as a cone or pyramid or even an iceberg. Just the tip, in this model, represents the knowledge we are both certain of and can explicitly represent. Going towards the base we find more and more knowledge about which we are less and less certain. For example, the central portion represents what can be obtained merely by term association. While this is a weaker form that the fully explicit tip it still is capable of representing significant portions of knowledge as Schvaneveldt has shown. Finally, near the base of the pyramid is a large body of very weak knowledge (mere recognition of the domain, for example) which are also used by human cognitive processes, and which we need to construct systems which also can make use of such weak, under-specified vague and ambiguous knowledge. Such a program of research would make significant inroads to the challenges currently facing the SW and AI more generally.

In this paper, we have touched on three views of what the SW is. There is also a fourth view, much harder to define and discuss, which

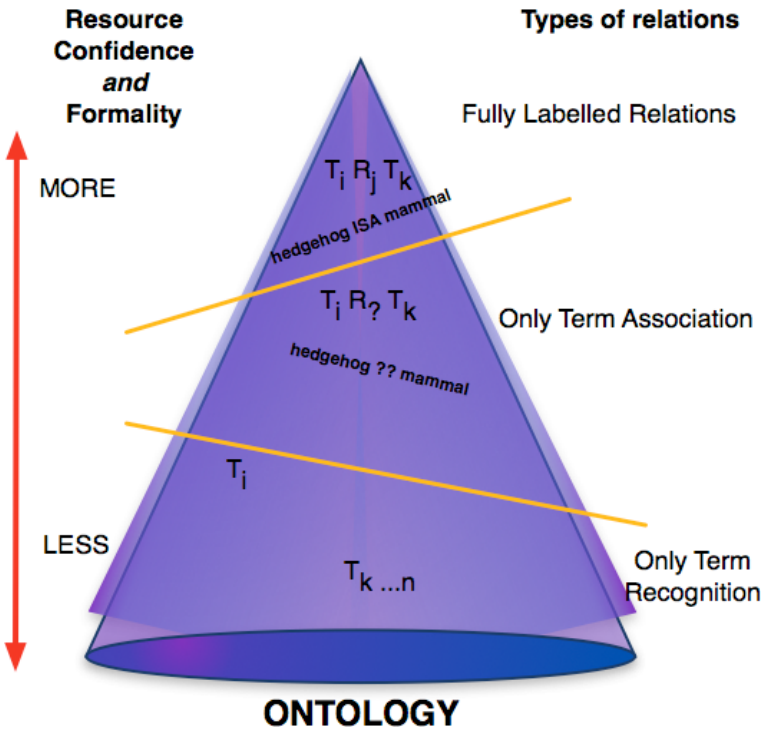

Fig. 5.1 Ways of presenting knowledge.

is that if the SW just keeps moving as an engineering development and is lucky (as the successful scale-up of the WWW seems to have been luckier, or better designed, than many cynics expected) then real problems will not arise. This view is a hunch and not open to close analysis but one can only wish it well, as it were, without being able to discuss it in detail further at this stage. It remains the case that the SW has not yet taken off, as the WWW did, and Google-IR and iPods did; it may be that something about its semantics is holding it back, and that may be connected, as we have argued, to its ability to generate semi-formalised material on a great scale from existing WWW material.

The main argument of the paper has been that NLP will continue to underlie the SW, including its initial construction from unstructured sources like the WWW, in several different ways, and whether it advocates realise this or not: chiefly, we argued, such NLP activity is the

only way up to a defensible notion of meaning at conceptual levels (in the original SW diagram) based on lower-level empirical computations over usage. Our aim is definitely not to claim logic-bad, NLP-good in any simple-minded way, but to argue that the SW will be a fascinating interaction of these two methodologies, again like the WWW (which has been basically a field for statistical NLP research) but with deeper content. Only NLP technologies (and chiefly IE) will be able to provide the requisite RDF knowledge stores for the SW from existing WWW (unstructured) text data bases, and in the vast quantities needed. There is no alternative at this point, since a wholly or mostly hand-crafted SW is also unthinkable, as is a SW built from scratch and without reference to the WWW. We also assume that, whatever the limitations on current SW representational power we have drawn attention to here, the SW will continue to grow in a distributed manner so as to serve the needs of scientists, even if it is not perfect. The WWW has already shown how an imperfect artefact can become indispensable.

Contemporary statistical large-scale NLP offers new ways of looking at usage in detail and in quantity — even if the huge quantities required now show we cannot easily relate them to an underlying theory of human learning and understanding. We can see glimmerings, in machine learning studies, of something like Wittgenstein's [197] language games' in action, and of the role of key concepts in the representation of a whole language. Part of this can only be done by some automated recapitulation of the role primitive concepts play in the organisation of (human-built) ontologies, thesauri, and wordnets. The heart of the issue is the creation of meaning by some interaction of (unstructured language) usage and the interpretations to be given to higher-level concepts — this is a general issue, but the construction of the SW faces it crucially and it could be the critical arena for progress on a problem that goes back at least to Kant's classic formulation in terms of "concepts without percepts are empty, percepts without concepts are blind". If we see that opposition as one of language data (like percepts) to concepts, the risk is of formally defined concepts always remaining empty (see discussions of SW meaning in Horrocks and Patel-Schneider [83]). The answer is, of course, to find a way, upwards, from one to the other.

Acknowledgments

CB would like to thank José Iria for help in discussing and formulating the Abraxas model, and Ziqi Zhang for undertaking parts of the implementation and evaluation. Our thanks also to Kieron O'Hara and Wendy Hall for comments, discussion and aid with editing. CB has been supported in this work by the EPSRC project Abraxas (http://nlp.shef.ac.uk/abraxas/) under grant number GR/T22902/01 and the EC funded IP Companions IST-034434 (www.companions-project.org)

References

[1] K. Ahmad, "Pragmatics of specialist terms: The acquisition and representation of terminology," in *Proceedings of the Third International EAMT Workshop on Machine Translation and the Lexicon*, pp. 51–76, London, UK: Springer-Verlag, 1995.

[2] K. Ahmad, "Terminology and artificial intelligence," in *The Handbook of Terminology Management*, (S.-E. Wright and G. Budin, eds.), Amsterdam, Philadelphia: John Benjamins, 2000.

[3] D. Artz and Y. Gil, "A survey of trust in computer science and the semantic web," submitted for publication, 2006.

[4] S. Azzam, K. Humphreys, and R. Gaizauskas, "Using conference chains for text summarization," in *Proceedings of the ACL'99 Workshop on Conference and its Applications,* Baltimore, 1999.

[5] L. Ballesteros and W. B. Croft, "Resolving ambiguity for cross-language retrieval," in *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 64–71, New York, NY, USA: ACM Press, 1998.

[6] J. Bear, D. Israel, J. Petit, and D. Martin, "Using information extraction to improve document retrieval," in *The Sixth Text REtrieval Conference (TREC-6)*, pp. 367–377, National Institute of Standards and Technology (NIST), 1997.

[7] N. Bely, N. Siot-Deauville, and A. Borillo, *Procedures d'Analyse Semantique Appliquees a la Documentation Scientifique*. Paris: Gauthier-Villars, 1970.

[8] A. L. Berger and J. D. Lafferty, "Information retrieval as statistical translation," in *SIGIR*, pp. 222–229, ACM, 1999.

[9] H. Bergson, *Le rire: Essai sur la signification du comique*. Paris: Editions Alcan, 1900.

[10] T. Berners-Lee, W. Hall, J. A. Hendler, K. O'Hara, N. Shadbolt, and D. J. Weitzner, "A framework for web science," *Foundations and Trends in Web Science*, vol. 1, no. 1, pp. 1–130, 2006.

[11] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific American*, pp. 30–37, 2001.

[12] D. Bikel, S. Miller, R. Schwartz, and R. Weischedel, "Nymble: A high-performance learning name-finder," in *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP 97)*, 1997.

[13] D. G. Bobrow and T. Winograd, "On overview of krl, a knowledge representation language," *Cognitive Science*, vol. 1, no. 1, pp. 3–46, 1977.

[14] K. Bontcheva, A. Kiryakov, O. Lab, A. H. B. Blvd, H. Cunningham, B. Popov, and M. Dimitrov, "Semantic web enabled, open source language technology," in *Language Technology and the Semantic Web, Workshop on NLP and XML (NLPXML-2003)*, Held in conjunction with EACL 2003, Budapest, 2003.

[15] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman, "NYU: Description of the MENE named entity system as used in MUC-7," in *Message Understanding Conference Proceedings MUC-7*, 1998.

[16] R. Braithwaite, *Scientific Explanation*. Cambridge University Press, 1953.

[17] C. Brewster, "Mind the gap: Bridging from text to ontological knowledge," PhD thesis, Department of Computer Science, University of Sheffield, 2008.

[18] C. Brewster, H. Alani, S. Dasmahapatra, and Y. Wilks, "Data driven ontology evaluation," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC-04),* Lisbon, Portugal, 2004.

[19] C. Brewster, F. Ciravegna, and Y. Wilks, "Knowledge acquisition for knowledge management: Position paper," in *Proceedings of the IJCAI-2001 Workshop on Ontology Learning*, Seattle, WA: IJCAI, 2001.

[20] C. Brewster, F. Ciravegna, and Y. Wilks, "Background and foreground knowledge in dynamic ontology construction," in *Proceedings of the Semantic Web Workshop,* Toronto, August 2003, SIGIR, 2003.

[21] C. Brewster, J. Iria, Z. Zhang, F. Ciravegna, L. Guthrie, and Y. Wilks, "Dynamic iterative ontology learning," in *Recent Advances in Natural Language Processing (RANLP 07),* Borovets, Bulgaria, 2007.

[22] C. Brewster and Y. Wilks, "Ontologies, taxonomies, thesauri learning from texts," in *The Keyword Project: Unlocking Content through Computational Linguistics*, (M. Deegan, ed.), Centre for Computing in the Humanities, Kings College London. *Proceedings of the the Use of Computational Linguistics in the Extraction of Keyword Information from Digital Library Content: Workshop 5–6 February,* 2004.

[23] E. Brill, "Some advances in transformation-based part of speech tagging," in *Proceedings of the 12th National Conference on Artificial Intelligence*, pp. 722–727, Seattle, WA: AAAI Press, 1994.

[24] P. Brown, J. Cocke, S. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to machine translation," *Computational Linguistics*, vol. 16, pp. 79–85, 1990.

[25] P. F. Brown, J. Cocke, S. D. Pietra, V. J. D. Pietra, F. Jelinek, R. L. Mercer, and P. S. Roossin, "A statistical approach to language translation," in *Proceedings of the 12th International Conference on Computational*

Linguistics — COLING 88, pp. 71–76, Budapest, Hungary: John von Neumann Society for Computing Sciences, 1988.

[26] C. Cardie, "Empirical methods in information extraction," *AI Journal*, vol. 18, no. 4, pp. 65–79, 1997.

[27] R. Carnap, *Semantics and Necessity*. Chicago: University of Chicago Press, 1946.

[28] E. Charniak, "Jack and janet in search of a theory of knowledge," in *Proceedings of the 3rd International Joint Conference on Artificial Intelligence — IJCAI*, (N. Nilsson, ed.), pp. 337–343, Stanford, CA: William Kaufmann, 1973.

[29] C. Chelba and F. Jelinek, "Exploiting syntactic structures for language and modelling," in *Proceedings of ACL 98,* Montreal, Canada, 1998.

[30] Y. Chiaramella and J.-Y. Nie, "A retrieval model based on an extended modal logic and its application to the rime experimental approach," in *SIGIR'90, Proceedings of the 13th International Conference on Research and Development in Information Retrieval*, (J.-L. Vidick, ed.), pp. 25–43, Brussels, Belgium: ACM, 5–7 September, 1990.

[31] N. Chomsky, *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press, 1965.

[32] G. E. Churcher, E. S. Atwell, and C. Souter, "Generic template for the evaluation of dialogue management systems," in *Proceedings of Eurospeech '97*, pp. 2247–2250, 1997.

[33] P. Cimiano, A. Pivk, L. Schmidt-Thieme, and S. Staab, "Learning taxonomic relations from heterogeneous sources of evidence," in *Ontology Learning from Text: Methods, Evaluation and Applications, Frontiers in Artificial Intelligence*, IOS Press, 2005.

[34] F. Ciravegna, "$(LP)^2$, an adaptive algorithm for information extraction from web-related texts," in *Proceedings of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining held in conjunction with the 17th International Joint Conference on Artificial Intelligence*, 2001.

[35] F. Ciravegna, S. Chapman, A. Dingli, and Y. Wilks, "Learning to harvest information for the semantic web," in *ESWS, Lecture Notes in Computer Science*, (C. Bussler, J. Davies, D. Fensel, and R. Studer, eds.), pp. 312–326, Springer, 2004.

[36] F. Ciravegna and Y. Wilks, "Designing adaptive information extraction for the semantic web in amilcare," in *Annotation for the Semantic Web*, (S. Handschuh and S. Staab, eds.), Amsterdam: IOS Press, 2003.

[37] R. Collier, "Automatic template creation for information extraction," PhD thesis, University of Sheffield, Computer Science Dept., 1998.

[38] J. Cowie, L. Guthrie, W. Jin, W. Odgen, J. Pustejovsky, R. Wanf, T. Wakao, S. Waterman, and Y. Wilks, "Crl/brandeis: The diderot system," in *Proceedings of Tipster Text Program (Phase I)*, Morgan Kaufmann, 1993.

[39] W. B. Croft and J. Lafferty, eds., *Language Modeling for Information Retrieval*. Dordrecht: Kluwer Academic Publishers, 2003.

[40] H. Cunningham, "JAPE: A java annotation patterns engine," Research Memorandum CS-99-06, Department of Computer Science, University of Sheffield, 1999.

[41] H. Cunningham, K. Humphreys, R. Gaizauskas, and Y. Wilks, "GATE — a TIPSTER-based general architecture for text engineering," in *Proceedings of the TIPSTER Text Program Phase III*, Morgan Kaufmann, 1997.

[42] W. Daelemans, J. Zavrel, K. Sloot, and A. V. D. Bosch, "TiMBL: Tilburg memory-based learner, version 1.0, reference guide," Technical Report 98-03, Induction of Linguistic Knowledge, Tilburg University, 1998.

[43] C. V. Damme, M. Hepp, and K. Siorpaes, "Folksontology: An integrated approach for turning folksonomies into ontologies," in *Bridging the Gep between Semantic Web and Web 2.0 (SemNet 2007)*, pp. 57–70, 2007.

[44] G. F. DeJong, "Prediction and substantiation: A new approach to natural language processing," *Cognitive Science*, 1979.

[45] K. Dellschaft and S. Staab, "On how to perform a gold standard based evaluation of ontology learning," in *International Semantic Web Conference, Lecture Notes in Computer Science,* vol. 4273, (I. F. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, and L. Aroyo, eds.), pp. 228–241, Springer, 2006.

[46] A. Dingli, F. Ciravegna, D. Guthrie, and Y. Wilks, "Mining web sites using unsupervised adaptive information extraction," in *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics,* Budapest, Hungary, 2003.

[47] W. H. Dutton and E. J. Helsper, "The internet in britain: 2007," Technical report, Oxford Internet Institute, University of Oxford, Oxford, UK, 2007.

[48] O. Etzioni, M. J. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, "Unsupervised named-entity extraction from the web an experimental study," *Artificial Intelligence*, vol. 165, no. 1, pp. 91–134, 2005.

[49] D. Fass and Y. A. Wilks, "Preference semantics, ill-formedness, and metaphor," *American Journal of Computational Linguistics*, vol. 9, no. 3–4, pp. 178–187, 1983.

[50] J. Fodor and B. P. McLaughlin, "Connectionism and the problem of systematicity: Why smolensky's solution doesn't work," *Cognition*, vol. 35, no. 2, pp. 183–204, 1990.

[51] D. Gachot, E. Lage, and J. Yang, "The systran nlp browser: An application of mt technique in multilingual ir," in *Cross Language Information Retrieval*, (G. Greffenstette, ed.), 1998.

[52] R. Gaizauskas and Y. Wilks, "Information extraction: Beyond document retrieval," Technical report CS–97–10, Department of Computer Science, University of Sheffield, 1997.

[53] A. Gangemi, N. Guarino, and A. Oltramari, "Conceptual analysis of lexical taxonomies the case of WordNet top-level," in *Proceedings of the International Conference on Formal Ontology in Information Systems*, pp. 285–296, ACM Press, 2001.

[54] J.-C. Gardin "*SYNTOL*," Graduate School of Library Service, Rutgers, The State University, New Brunswick, NJ, 1965.

[55] R. Garside, "The CLAWS word-tagging system," in *The Computational Analysis of English: A Corpus-based Approach*, (R. Garside, G. Leech, and G. Sampson, eds.), pp. 30–41, London: Longman, 1987.

[56] J. J. Gibson, "What gives rise to the perception of motion?," *Psychological Review*, vol. 75, no. 4, pp. 335–346, 1968.

[57] C. F. Goldfarb, "Sgml: The reason why and the first published hint," *Journal of the American Society for Information Science*, vol. 48, no. 7, 1997.

[58] T. Gollins and M. Sanderson, "Improving cross language information retrieval with triangulated translation," in *SIGIR*, (W. B. Croft, D. J. Harper, D. H. Kraft, and J. Zobel, eds.), pp. 90–95, ACM, 2001.

[59] A. Gómez-Pérez, "Evaluation of ontologies," *International Journal of Intelligent Systems*, vol. 16, no. 3, pp. 391–409, 2001.

[60] R. Granger, "Foulup: A program that figures out meanings of words from context," in *Proccedings of the 5th International Joint Conference on Artificial Intelligence (IJCAI 77)*, 1977.

[61] B. F. Green, A. K. Wolf, C. Chomsky, and K. Laughery, "BASEBALL: An automatic question answerer," in *Readings in Natural Language Processing*, (B. J. Grosz and B. L. Webber, eds.), pp. 545–549, Los Altos: Morgan Kaufmann, 1961.

[62] G. Grefenstette, "The world wide web as a resource for example-based machine translation tasks," in *Translating and the Computer 21: Proceedings of the 21st International Conference on Translating and the Computer*, London: ASLIB, 1999.

[63] G. Grefenstette and M. A. Hearst, "Method for refining automatically-discovered lexical relations: Combining weak techniques for stronger results," in *Proceedings of the AAAI Workshop on Statistically-based Natural Language Processing Techniques*, AAAI Press, 1992.

[64] R. Grishman, "Information extraction techniques and challenges," in *Information Extraction: A Multidisciplinary Approach to an Emerging Technology*, (M. T. Pazienza, ed.), 1997.

[65] R. Grishman and J. Sterling, "Acquisition of selectional patterns," in *14th International Conference on Computational Linguistics — COLING*, pp. 658–664, Nantes, France, 1992.

[66] M. Gross, "On the equivalence of models of language used in the fields of mechanical translation and information retrieval," *Information Storage and Retrieval*, vol. 2, no. 1, pp. 43–57, 1964.

[67] T. Gruber, "Collective knowledge systems: Where the social web meets the semantic web," *Web Semant*, vol. 6, no. 1, pp. 4–13, 2008.

[68] N. Guarino, "Some ontological principles for designing upper level lexical resources," in *Proceedings of the First International Conference on Language Resources and Evaluation, LREC 98*, 1998.

[69] D. Guthrie, B. Allison, W. Liu, L. Guthrie, and Y. Wilks, "A closer look at skip-gram modelling," in *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC-2006),* Genoa, Italy, 2006.

[70] H. Halpin, "The semantic web: The origins of artificial intelligence redux," in *Third International Workshop on the History and Philosophy of Logic, Mathematics, and Computation (HPLMC-04 2005),* Donostia San Sebastian, Spain, 2004.

[71] S. M. Harabagiu, D. I. Moldovan, M. Pasca, M. Surdeanu, R. Mihalcea, R. Girju, V. Rus, V. F. Lacatusu, P. Morarescu, and R. C. Bunescu, "Answering complex, list and context questions with LCC's question-answering server," in *TREC*, 2001.

[72] D. Harman and D. Marcu, eds., *DUC 2001 Workshop on Text Summarization*. 2001.

[73] J. Haugeland, *Artificial Intelligence The Very Idea*. MIT Press, 1985.

[74] P. J. Hayes, "The naive physics manifesto," in *Expert Systems in the Electronic Age*, (D. Michie, ed.), pp. 242–270, Edinburgh, Scotland: Edinburgh University Press, 1979.

[75] M. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *Proceedings of the 14th International Conference on Computational Linguistics (COLING 92),* Nantes, France, July 1992.

[76] J. A. Hendler, "Knowledge is power: A view from the semantic web," *AI Magazine*, vol. 26, no. 4, pp. 76–84, 2005.

[77] E. Herder, *An Analysis of User Behaviour on the Web-Site*. VDM Verlag, 2007.

[78] C. Hewitt, "Procedural semantics," in *Natural Language Processing*, (Rustin, ed.), New York: Algorithmics Press, 1973.

[79] G. Hirst, "Context as a spurious concept," in *Conference on Intelligent Text Processing and Computational Linguistics*, pp. 273–287, Mexico City, 2000.

[80] J. Hobbs, "The generic information extraction system," in *Proceedings of the 5th Message Understanding Conference (MUC)*, 1993.

[81] J. R. Hobbs, "Ontological promiscuity," in *Proceedings of 23rd Annual Meeting of the Association for Computational Linguistics–ACL 85*, pp. 61–69, University of Chicago, Chicago, Illinois, USA, 1985.

[82] I. Horrocks, "Description logics in ontology applications," in *KI 2005: Advances in Artificial Intelligence, 28th Annual German Conference on AI, Lecture Notes in Computer Science*, (U. Furbach, ed.), p. 16, Springer, vol. 3698, 2005.

[83] I. Horrocks and P. F. Patel-Schneider, "Three theses of representation in the semantic web," in *WWW '03: Proceedings of the 12th International Conference on World Wide Web*, pp. 39–47, New York, NY, USA: ACM, 2003.

[84] I. Horrocks, P. F. Patel-Schneider, and F. van Harmelen, "Reviewing the design of daml+oil: An ontology language for the semantic web," in *AAAI/IAAI*, pp. 792–797, 2002.

[85] I. R. Horrocks, "Using an expressive description logic: FaCT or fiction?," in *Proceedings of the 6th International Conference on Principles of Knowledge Representation and Reasoning (KR-98)*, (A. G. Cohn, L. Schubert, and S. C. Shapiro, eds.), pp. 636–649, San Francisco: Morgan Kaufmann Publishers, 1998.

[86] E. H. Hovy, "Key toward large-scale shallow semantics for higher-quality nlp," in *Proceedings of the 12th PACLING Conference,* Tokyo, Japan, 2005.

[87] W. J. Hutchins, "Linguistic processes in the indexing and retrieval of documents," *Linguistics*, 1971.

[88] K. Jeffrey, "What's next in databases?," *ERCIM News*, vol. 39, pp. 24–26, 1999.
[89] F. Jelinek and J. Lafferty, "Computation of the probability of initial substring generation by stochastic context free grammars," *Computational Linguistics*, vol. 17, no. 3, pp. 315–323, 1991.
[90] A. Kao and S. R. Poteet, eds., *Natural Language Processing and Text Mining*. Springer-Verlag, New York, Inc. Secaucus, NJ, USA, 2007.
[91] J. J. Katz, *Semantic Theory*. New York: Harper and Row Publishers, 1972.
[92] A. Kilgariff and G. Greffenstete, "Introduction to special issue of computational linguistics on: The web as corpus," *Computational Linguistics*, 2001.
[93] R. Krovetz, "More than one sense per discourse," in *Proceedings of the ACL-SIGLEX Workshop*, ACL, 1998.
[94] T. Kuhn, *The Structure of Scientific Revolutions*. Chicago, Illinois: University of Chicago Press, 1962.
[95] G. Leech, R. Garside, and M. Bryant, "Claws4: The tagging of the british national corpus," in *COLING — 15th International Conference on Computational Linguistics*, pp. 622–628, 1994.
[96] W. Lehnert, C. Cardie, D. Fisher, J. McCarthy, and E. Riloff, "University of massachusetts: Description of the circus system as used for muc-4," in *Proceedings of the 4th Message Understanding Conference MUC-4*, pp. 282–288, Morgan Kaufmann, 1992.
[97] W. G. Lehnert, "A conceptual theory of question answering," in *Proceedings of the 5th International Joint Conference on Artificial Intelligence – IJCAI 77*, (R. Reddy, ed.), pp. 158–164, William Kaufmann, 1977.
[98] W. G. Lehnert, *The Process of Question Answering: A Computer Simulation of Cognition*. Hillsdale, NJ: Lawrence Erlbaum Associates, Based on 1977 Yale PhD thesis, 1978.
[99] D. Lenat, M. Prakash, and M. Shepherd, "CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks," *AI Magazine*, vol. 6, no. 4, pp. 65–85, 1986.
[100] D. B. Lenat, "Cyc: A large-scale investment in knowledge infrastructure," *Communications of the ACM*, vol. 38, no. 11, pp. 33–38, 1995.
[101] D. B. Lenat and R. V. Guha, *Building Large Knowledge-Based Systems*. Reading, MA: Addison-Wesley, 1990.
[102] D. Lewis, "General semantics," in *The Semantics of Natural Language*, (D. Davidson and Harman, eds.), Amsterdam: Kluwer, 1972.
[103] S. F. Liang, S. Devlin, and J. Tait, "Using query term order for result summarisation," in *SIGIR*, (R. A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, and J. Tait, eds.), pp. 629–630, ACM, 2005.
[104] W. Lin, R. Yangarber, and R. Grishman, "Bootstrapped learning of semantic classes from positive and negative examples," in *Proceedings of the 20th International Conference on Machine Learning: ICML 2003 Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining,* Washington, DC, 2003.
[105] H. Longuet-Higgins, "The algorithmic description of natural language," *Proceedings of the Royal Society of London B*, vol. 182, pp. 255–276, 1972.

[106] J. S. Luciano and R. D. Stevens, "e-Science and biological pathway semantics," *BMC Bioinformatics*, vol. 8, Suppl. 3, p. S3, 2007.

[107] H. P. Luhn, "A statistical approach to the mechanized encoding and searching of literary information," *IBM Journal of Research and Development*, vol. 1, no. 4, pp. 309–317, 1957.

[108] H. P. Luhn, "Automatic creation of literature abstracts," *IBM Journal of Research and Development*, vol. 2, pp. 159–165, 1958.

[109] K. Mahesh, S. Nirenburg, and S. Beale, "Toward full-text ontology-based word sense disambiguation," in *Recent Advances in Natural Language Processing II*, (N. Nicolov and R. Mitkov, eds.), Amsterdam: John Benjamins, 2000.

[110] I. Mani, *Automatic Summarization*. Amsterdam: J. Benjamins Pub. Co., 2001.

[111] D. Marr, *Vision: A Computational Investigation in the Human Representation and Processing of Visual Information*. San Francisco: W. H. Freeman, 1981.

[112] C. C. Marshall and F. M. Shipman, "Which semantic web?," in *HYPERTEXT '03: Proceedings of the 14th ACM Conference on Hypertext and Hypermedia*, pp. 57–66, New York, NY, USA: ACM Press, 2003.

[113] M. Masterman and Y. A. Wilks, eds., *Language, Cohesion and Form*. Studies in Natural Language Processing. Cambridge University Press, 2005.

[114] M. L. Mauldin, "Retrieval performance in ferret a conceptual information retrieval system," in *SIGIR '91: Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 347–355, New York, NY, USA: ACM Press, 1991.

[115] D. Maynard and S. Ananiadou, "TRUCKS a model for automatic term recognition," *Journal of Natural Language Processing*, 2000.

[116] J. McCarthy and P. J. Hayes, "Some philosophical problems from the standpoint of artificial intelligence," in *Machine Intelligence 4*, (B. Meltzer, D. Michie, and M. Swann, eds.), pp. 463–502, Edinburgh, Scotland: Edinburgh University Press, 1969.

[117] J. McCarthy and Lifschitz, *Formalization of Common Sense*. Norwood, NJ: Ablex, 1990.

[118] D. McDermott, "Artificial intelligence meets natural stupidity," in *Mind Design: Philosophy, Psychology, Artificial Intelligence*, (J. Haugeland, ed.), pp. 143–160, Cambridge, MA: The MIT Press, 1981.

[119] D. H. Mellor, "Natural kinds," *Brtish Journal of Philosoptical Science*, vol. 28, no. 4, pp. 299–312, 1977.

[120] G. Miller, "Wordnet: A lexical database for english," *Communications of ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[121] G. A. Miller, "WordNet: An online lexical database," *International Journal of Lexicography*, 1990.

[122] R. K. Moore, "A comparison of data requirements for asr systems and human listeners," in *Proceedings of EUROSPEECH 2003*, 2003.

[123] J. Morato, M. N. Marzal, J. Llorns, and J. Moreiro, "Wordnet applications," in *Proceedings of the 2nd Global Wordnet Conference*, Brno. Czsech Republic, 2004.

[124] A. A. Morgan, L. Hirschman, M. Colosimo, A. S. Yeh, and J. B. Colombe, "Gene name identification and normalization using a model organism database," *J Biomedical Information*, vol. 37, no. 6, pp. 396–410, 2004.

[125] R. Morgan, R. Garigliano, P. Callaghan, S. Poria, M. Smith, A. Urbanowicz, R. Collingham, M. Costantino, C. Cooper, and T. L. Group, "University of durham: Description of the LOLITA system as used in MUC-6," in *Proceedings of the Message Understanding Conference MUC6*, Morgan Kaufmann, 1995.

[126] S. Muggleton, "Recent advances in inductive logic programming," in *Proceedings of the 7th Annual ACM Conference on Computional Learning Theory*, pp. 3–11, New York, NY: ACM Press, 1994.

[127] T. Nelson, *The Future of Information*. Tokyo: ASCII Press, 1997.

[128] S. Nirenburg, "Pangloss: A machine translation project," in *HLT*, Morgan Kaufmann, 1994.

[129] S. Nirenburg, R. E. Frederking, D. Farwell, and Y. Wilks, "Two types of adaptive MT environments," in *15th International Conference on Computational Linguistics – COLING 94*, pp. 125–128, Kyoto, Japan, 1994.

[130] S. Nirenburg, H. Somers, and Y. Wilks, eds., *Readings in Machine Translation*. Cambridge, MA: MIT Press, 2003.

[131] S. Nirenburg and Y. Wilks, "What's in a symbol ontology, representation, and language," *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 13, no. 1, pp. 9–23, 2001.

[132] B. Norton, S. Foster, and A. Hughes, "A compositional operational semantics for OWL-S," in *Proceedings of the Formal Techniques for Computer Systems and Business Processes, European Performance Engineering Workshop, EPEW 2005 and International Workshop on Web Services and Formal Methods, WS-FM 2005,* Versailles, France, September 1–3, 2005, Lecture Notes in Computer Science, (M. Bravetti, L. Kloul, and G. Zavattaro, eds.), pp. 303–317, vol. 3670, Springer, 2005.

[133] J. Oh, Y. Chae, and K. Choi, "Japanese term extraction using a dictionary hierarchy and an MT system," *Terminology*, 2000.

[134] K. O'Hara, H. Alani, Y. Kalfoglou, and N. Shadbolt, "Trust strategies for the semantic web," in *ISWC Workshop on Trust, Security, and Reputation on the Semantic Web, CEUR Workshop Proceedings,* CEUR-WS.org, vol. 127, (J. Golbeck, P. A. Bonatti, W. Nejdl, D. Olmedilla, and M. Winslett, eds.), 2004.

[135] J. Olney, C. Revard, and P. Ziff, *Some Monsters in Noah's Ark*. Research Memorandum, Systems Development Corp., Santa Monica, CA, 1968.

[136] A. Oltramari, A. Gangemi, N. Guarino, and C. Masolo, "Restructuring wordnet's top-level: The ontoclean approach," in *Proceedings of the Workshop OntoLex'2, Ontologies and Lexical Knowledge Bases*, 2002.

[137] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking bringing order to the web," Technical Report, Stanford University, 1999.

[138] P. F. Patel-Schneider, P. Hayes, and I. Horrocks, "Owl web ontology language semantics and abstract syntax," Technical Report, W3C Recommendation, 2004.

[139] J. Pearl, "Bayesian networks: A model of self-activated memory for evidential reasoning," in *Proceedings of Cognitive Science Society (CSS-7)*, 1985.

[140] W. Peters and Y. Wilks, "Distribution-oriented extension of wordnet's ontological framework," in *Proceedings of Recent Advances in Natural language Processing*, Tzigov Chark, Bulgaria, 2001.

[141] E. Pietrosanti and B. Graziadio, "Extracting information for business needs," Unicom Seminar on Information Extraction, London, 1997.

[142] J. B. Pollack, "Recursive distributed representations," *Artificial Intelligence*, vol. 46, no. 1–2, pp. 77–105, 1990.

[143] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 275–281, New York, NY, USA: ACM Press, 1998.

[144] K. R. Popper, *The Logic of Scientific Discovery*. London: Hutchinson, 1959.

[145] J. Pustejovsky, *The Generative Lexicon*. Cambridge, MA: The MIT Press, 1995.

[146] J. Pustejovsky, "Type construction and the logic of concepts," in *The Language of Word Meaning*, (F. Busa and P. Bouillon, eds.), Cambridge University Press, 2001.

[147] H. Putnam, "Is semantics possible?," *Metaphilosophy*, vol. 1, no. 3, pp. 187–201, 1970. (Reprinted in H. Putnam, 1985).

[148] H. Putnam, "The meaning of 'meaning'," in *Language, Mind, and Knowledge. Minnesota Studies in the Philosophy of Science, vol. 7*, (K. Gunderson, ed.), pp. 131–193, Minneapolis, Minnesota: University of Minnesota Press, 1975. Reprinted in H. Putnam, (1985).

[149] H. Putnam, *Philosophical Papers, vol. 2: Mind, Language and Reality*. Cambridge: Cambridge University Press, 1985.

[150] W. V. Quine, "Two dogmas of empiricism," *The Philosophical Review*, vol. 60, pp. 20–43, 1951. (Reprinted in W.V.O. Quine, From a Logical Point of View Harvard University Press, 1953; second, revised, edition 1961).

[151] P. Resnik, "Selectional constraints: An information-theoretic model and its computational realization," *Cognition*, vol. 61, pp. 127–159, 1996.

[152] E. Riloff and W. Lehnert, "Automated dictionary construction for information extraction from text," in *Proceedings of the 9th Conference on Artificial Intelligence for Applications (CAIA'93)*, pp. 93–99, Los Alamitos, CA, USA: IEEE Computer Society Press, 1993.

[153] E. Riloff and J. Shoen, "Automatically acquiring conceptual patterns without an automated corpus," in *Proceedings of the 3rd Workshop on Very Large Corpora*, (D. Yarovsky and K. Church, eds.), pp. 148–161, 1995. Association for Computational Linguistics, Somerset, New Jersey: Association for Computational Linguistics.

[154] E. Roche and Y. Schabes, "Deterministic part-of-speech tagging with finite-state transducers," *Computational Linguistics*, vol. 21, no. 2, pp. 227–253, 1995.

[155] P. M. Roget, *Thesaurus of English Words and Phrases Classified and Arranged so as to Facilitate the Expression of Ideas and Assist in Literary composition*. London: Longman, Brown, Green, Longmans, 1852.

[156] N. Sager, *Natural Language Information Processing: A Computer Grammar of English and Its Applications*. Reading, MA: Addison-Wesley, 1981.

[157] H. Saggion and R. Gaizauskas, "Mining on-line sources for definition knowledge," in *Proceedings of the 17th International FLAIRS Conference*, pp. 45–52, Miami Beach, 2004.

[158] G. Salton, "A new comparison between conventional indexing (medlars) and automatic text processing (smart)," *Journal of the American Society for Information Science*, vol. 23, pp. 75–84, 1972.

[159] K. Samuel, S. Carberry, and K. Vijay-Shanker, "Dialogue act tagging with transformation-based learning," in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics – COLING-ACL 98*, pp. 1150–1156, Université de Montréal Montreal, Quebec, Canada, 1998.

[160] R. Schank, "Conceptual dependency: A theory of natural language understanding," *Cognitive Psychology*, 1972.

[161] R. W. Schvaneveldt, *Pathfinder Associative Networks Studies in Knowledge Organization*. Norwood, NJ: Ablex, 1990.

[162] E. Sirin and B. Parsia, "Pellet: An owl DL reasoner," in *Description Logics, CEUR Workshop Proceedings,* CEUR-WS.org, vol. 104, (V. Haarslev and R. Möller, eds.), 2004.

[163] B. Smith, "Formal ontology, common sense and cognitive science," *International Journal of Human-Computational Studies*, vol. 43, no. 5-6, pp. 641–667, 1995.

[164] B. Smith, J. Williams, and S. Schulze-Kremer, "The ontology of the gene ontology," *AMIA Annual Symposium Proceedings*, pp. 609–613, 2003.

[165] J. Sowa, "The challenge of knowledge soup," in *Proceedings of the 1st International WordNet Conference*, Goa, India, 2004.

[166] K. Spärck Jones, "Synonymy and semantic classification," PhD thesis, University of Cambridge, England. Published in the Edinburgh Information Technology Series (EDITS), S. Michaelson and Y. Wilks (eds.), Edinburgh, Scotland: Edinburgh University Press, 1986, 1964.

[167] K. Spärck Jones, *Retrieving Information or Answering Questions*. London: The British Library, 1990. The British Library Annual Research Lecture, 1989.

[168] K. Spärck Jones, "Information retrieval and artificial intelligence," *Artificial Intelligence*, vol. 114, no. 1–2, pp. 257–281, 1999.

[169] K. Spärck Jones, "What is the role of nlp in text retrieval?," in *Natural Language Information Retrieval*, (T. Strzalkowski, ed.), Dordrecht: Kluwer Academic Publishers, 1999.

[170] K. Spärck Jones, "What's new about the semantic web? some questions," *ACM SIGIR Forum*, vol. 38, pp. 18–23, 2004.

[171] K. Spärck Jones and J. R. Galliers, "Evaluating natural language processing systems: An analysis and review," *Machine Translation*, vol. 12, pp. 375–379, 1997. (Lecture Notes in Artificial Intelligence, 1083).

[172] R. Stevens, M. E. Aranguren, K. Wolstencroft, U. Sattler, N. Drummond, M. Horridge, and A. Rector, "Using owl to model biological knowledge," *International Journal of Human-Computational Studies*, vol. 65, no. 7, pp. 583–594, 2007.

[173] R. Stevens, O. Bodenreider, and Y. A. Lussier, "Semantic webs for life sciences: Session introduction," in *Pacific Symposium on Biocomputing-PSB06*, (R. B. Altman, T. Murray, T. E. Klein, A. K. Dunker, and L. Hunter, eds.), pp. 112–115, World Scientific, 2006.

[174] M. Stevenson and Y. Wilks, "Combining weak knowledge sources for sense disambiguation," in *IJCAI*, (T. Dean, ed.), pp. 884–889, Morgan Kaufmann, 1999.

[175] T. Strzalkowski and B. Vauthey, "Natural language processing in automated information retrieval," Proteus Project Memorandum, Department of Computer Science, New York University, 1991.

[176] T. Strzalkowski and J. Wang, "A self-learning universal concept spotter," in *Proceedings of the 16th International Conference on Computational Linguistics — COLING 96*, pp. 931–936, Center for Sprogteknologi, Copenhagen, Denmark, 1996.

[177] J. Surowiecki, *The Wisdom of Crowds*. New York: Random House, 2004.

[178] P. Velardi, M. Missikoff, and R. Basili, "Identification of relevant terms to support the construction of domain ontologies," in *Proceedings of the ACL 2001 Workshop on Human Language Technology and Knowledge Management*, Toulouse, 2001.

[179] M. Vilain, "Validation of terminological inference in an information extraction task," in *Proceedings of the ARPA Human Language Technology Workshop 93*, pp. 150–156, Princeton, NJ, (distributed as *Human Language Technology* by San Mateo, CA: Morgan Kaufmann Publishers), 1994.

[180] P. Vossen, "Introduction to eurowordnet," *Computers and the Humanities*, vol. 32, no. 2–3, pp. 73–89, 1998.

[181] D. Vrandecic, M. C. Suárez-Figueroa, A. Gangemi, and Y. Sure, eds., *Evaluation of Ontologies for the Web: 4th International EON Workshop, Located at the 15th International World Wide Web Conference WWW 2006*. 2006.

[182] J. Wilkins, *An Essay towards a Real Character and a Philosophical Language*. 1668. (Reprinted by Chicago University Press, 2002).

[183] Y. Wilks, "Text searching with templates," Cambridge Language Research Unit Memo ML 156, Cambridge University, 1964.

[184] Y. Wilks, "The application of CLRU's method of semantic analysis to information retrieval," Cambridge Language Research Unit Memo ML, 173, Cambridge University, 1965.

[185] Y. Wilks, "Computable semantic derivations," Technical Report SP-3017, Systems Development Corporation, 1968.

[186] Y. Wilks, "Knowledge structures and language boundaries," in *Proceedings of the 5th International Joint Conference on Artificial Intelligence — IJCAI*, (R. Reddy, ed.), pp. 151–157, Cambridge, MA: William Kaufmann, 1977.

[187] Y. Wilks, "Frames, semantics and novelty," in *Frame Conceptions and Text Understanding*, (Metzing, ed.), Berlin: de Gruyter, 1979.

[188] Y. Wilks, "Stone soup and the French room," in *Current Issues in Computational Linguistics: Essays in Honour of Don Walker*, (A. Zampolli, N. Calzolari, and M. Palmer, eds.), pp. 585–595, Pisa, Dordrecht: Giardini Editori e Stampatori, Kluwer Academic Publishers, 1994.

[189] Y. Wilks, "Companions: A new paradigm for agents," in *Proceedings of the International AMI Workshop, IDIAP*, Martigny, CH, 2004.

[190] Y. Wilks, "COMPUTER SCIENCE: Is there progress on talking sensibly to machines?," *Science*, vol. 318, no. 5852, pp. 927–928, 2007.

[191] Y. Wilks and R. Catizone, "Can we make information extraction more adaptive?," in *Information Extraction: Towards Scalable, Adaptable Systems, Lecture Notes in Artificial Intelligence*, (M. T. Pazienza, ed.), Springer Verlag, vol. 1714, 1999.

[192] Y. Wilks and J. Cowie, *Handbook of Natural Language Processing*. Marcel Dekker, "Information Extraction," pp. 241–260, 2000.

[193] Y. Wilks, B. M. Slator, and L. M. Guthrie, *Electric Words Dictionaries, Computers and Meaning*. Cambridge, MA: MIT Press, 1996.

[194] Y. A. Wilks and J. I. Tait, "A retrospective view of syonymy and semantic classification," in *Charting a New Course: Natural Language Processing and Information Retrieval: Essays in Honour of Karen Spärck Jones*, (J. I. Tait, ed.), Springer, 2005.

[195] T. Winograd, *Understanding Natural Language*. New York: Academic Press, 1972.

[196] T. Winograd and A. Flores, *Understanding Computers and Cognition: A New Foundation for Design*. Norwood, NJ: Ablex, 1986.

[197] L. Wittgenstein, *Philosophical Investigations*. Oxford: Blackwells, 1953.

[198] W. Woods, R. Kaplan, and B. Nash-Webber, "The lunar sciences natural language information system," Final Report 2378, Bolt, Beranek and Newman, Inc., Cambridge, MA, 1974.

[199] W. A. Woods, "What's in a link: Foundations for semantic networks," in *Representation and understanding: Studies in Cognitive Science*, (D. Bobrow and A. Collins, eds.), New York, NY: Academic Press, 1975.

[200] S.-H. Wu and W.-L. Hsu, "SOAT a semi-automatic domain ontology acquisition tool from chinese corpus," in *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, Academia Sinica, ACLCLP, and National Tsing Hua University, Taiwan, Morgan Kaufmann, 2002.

[201] D. Yarowsky, "Word-sense disambiguation using satistical models of roget's categories trained on large corpora," in *Proceedings of COLING-92*, pp. 454–460, Nantes, France, 1992.

[202] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *ACL – 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 189–196, Cambridge, MA, USA: Morgan Kaufmann, MIT, 1995.