Overview

Automatic construction of lexicons, taxonomies, ontologies, and other knowledge structures

Olena Medelyan,[1]* Ian H. Witten,[2] Anna Divoli[1] and Jeen Broekstra[3]

Abstract, structured, representations of knowledge such as lexicons, taxonomies, and ontologies have proven to be powerful resources not only for the systematization of knowledge in general, but to support practical technologies of document organization, information retrieval, natural language understanding, and question-answering systems. These resources are extremely time consuming for people to create and maintain, yet demand for them is growing, particularly in specialized areas ranging from legacy documents of large enterprises to rapidly changing domains such as current affairs and celebrity news. Consequently, researchers are investigating methods of creating such structures automatically from document collections, calling on the proliferation of interlinked resources already available on the web for background knowledge and general information about the world. This review surveys what is possible, and also outlines current research directions. © 2013 Wiley Periodicals, Inc.

How to cite this article:
WIREs Data Mining Knowl Discov 2013, 3: 257–279 doi: 10.1002/widm.1097

INTRODUCTION

Since time immemorial, people have striven to systematically represent their understanding of the world. With the advent of computers, abstract representations of knowledge can be operationalized and put to work. Encoding world knowledge in machine-readable form opens up new applications and capabilities. Statistically constructed dictionaries produce rough but useful machine translations; both manually and automatically constructed taxonomies generate effective metadata for finding documents; assertions are automatically acquired from the Web and assimilated into ontologies that are so accurate that algorithms can outperform people in answering complex questions.

Knowledge structures encode semantics in a way that is appropriate for the task they are intended to serve. They differ in coverage and depth, ranging from purpose-built resources for particular document collections, through domain-specific representations of varying depth, to extended efforts to capture comprehensive world knowledge in fine detail. Techniques for constructing lexicons, taxonomies, and ontologies automatically from documents and general web resources allow custom knowledge structures to be built for particular purposes. Improvements in accuracy and coverage underpin solutions to increasingly complex tasks. The world's richly connected nature is gradually becoming reflected in the World Wide Web itself, linking disparate knowledge structures so that they can benefit from each other's capabilities. With more knowledge, computers are getting smarter.

The automatic construction of knowledge structures draws on a range of disciplines, including knowledge engineering, information architecture, text mining, information retrieval, natural language processing, information extraction, and machine learning. This paper surveys the techniques that have been developed. We begin by introducing some of the key terms and concepts, an ontology of the

The authors have declared no conflicts of interest in relation to this article.

*Correspondence to: medelyan@gmail.com
[1]Pingar Research, Auckland, New Zealand
[2]University of Waikato, Hamilton, New Zealand
[3]Rivuli Development, Wellington, New Zealand

DOI: 10.1002/widm.1097

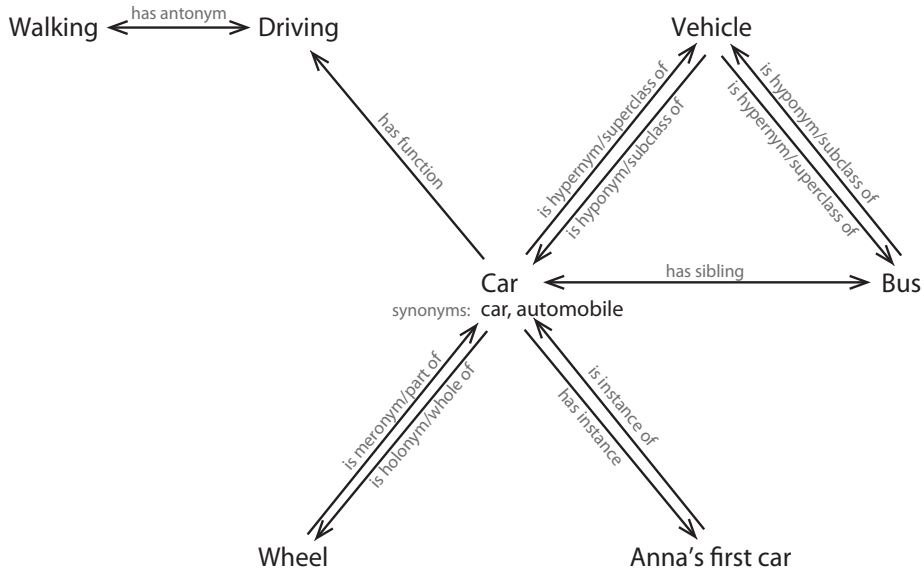

FIGURE 1 | Examples of semantic relations.

ontological domain—calling to mind the Ouroboros, an ancient symbol depicting a serpent or dragon eating its own tail that finds echoes in M.C. Escher's recursively space-filling tessellations of lizards. Following that, we briefly survey existing taxonomies, ontologies, and other knowledge structures before examining the various stages involved in mining meaning from text: identification of terms, disambiguation of referents, and extraction of relationships. We discuss various techniques that have been developed to assist in the automatic inference of knowledge structures from text, and the use of pre-existing knowledge sources to enrich the representation. We turn next to the key question of evaluating the accuracy of the knowledge structures that are produced, before identifying some trends in the research literature. Finally, we draw some\conclusions.

FROM WORDS TO KNOWLEDGE REPRESENTATION

Ontology is commonly described as the study of the nature of things, and an *ontology* is a means of organizing and conceptualizing a domain of interest. We use the term 'knowledge structure' to embrace dictionaries and lexicons, taxonomies, and full-blown ontologies, in order of increasing power and depth. This section introduces these concepts, along with some supporting terms.

Semantics of Language and Knowledge

The overall goal of knowledge structures is to encode semantics. The smallest unit of language that carries semantics is the *morpheme*. Morphemes may be free or bound. The former are independent *words* like *school* or *home*. The latter are attached to other words to modify their meaning: *-ing* generates the word *schooling* and *-less* the word *homeless*. In some cases, two standalone words are joined into a new word like *homeschooling*, or into multiword phrases, also called compound words, like *school bus* or *rest home*. *Concepts* typically represent classes of things, entities, or ideas, whose individual members are called *instances*. *Terms* are words or phrases that denote, or name, concepts. Figure 1 shows concepts such as CAR (with a further term adding the denotation *automobile*), WHEEL and VEHICLE, as well as one instance, ANNA'S FIRST CAR. In general, the relations between semantic units such as morphemes, words, terms, and concepts are called *semantic relations*.

If a term denotes more than one concept, which happens when a word has homonyms or is polysemous, the issue of ambiguity arises. Both *homonymy* and *polysemy* concern the use of the same word to express different meanings. In homonymy, the meanings are distinct (*bank* as a financial institution or the side of a river); in polysemy they are subsenses of the word (*bank* as a financial institution and *bank* as a building where such institution offers services). It is the context in which a word is used that helps us decode its intended meaning. For example, the word *house* in the context of oligarchy or government is likely to denote the concept *Royal Dynasty*.

It is often the case that more than one term can denote a given concept. For example, both *vocalist* and *singer* denote the concept *Singer*, or 'a person who

sings'. The semantic relation between these two terms is called synonymy; it expresses equivalence of meaning (e.g., *automobile* and *car* are equivalent terms that both denote the concept *Car* in Figure 1). The opposite relation is *antonymy* (*hot* and *cold*; *Walking* and *Driving* in Figure 1).

Semantic units relate to each other hierarchically when the meaning of one is *broader* or *narrower* than the meaning of the other. A specific type of hierarchical relation occurs between two concepts when one class of things subsumes the other. For example, *Singer* subsumes *Pop Singer* and *Opera Singer*, whereas *Vehicle* subsumes *Car*—in other words, *Vehicle* is a *hypernym* of *Car*. Another type of hierarchical relation is one between a concept and an instance of it, e.g., *Alicia Keys is-an-instance-of Pop Singer*. One concept can also be narrower than another because it denotes a particular part of it, e.g., *Wheel* is a part of *Car* in Figure 1; in other words, a *meronym*.

There are also many nonhierarchical relations, which can be grouped generically as 'a concept is related to another concept' (*Singer* has-related *Band*) or characterized more specifically (*Singer* is-member-of *Band* and *Singer* is-performing *Songs*).

Although the terminology outlined above is standard in linguistics, publishers of knowledge sources do not always use it consistently. For example, the word *term* in the context of taxonomies is typically used to mean *Concept*, and the word *label* in a taxonomy, which occurs in phrases such as *preferred* and *alternative labels* to denote different kinds of synonym, is used as the sense of *Term* as defined in this section.

Types of Knowledge Structure

Knowledge structures differ markedly in their specificity and the expressiveness of the meaning they encode. Some capture only basic knowledge such as the terms used in a particular domain, and their synonyms. Others encode a great deal more information about different concepts, the terms that denote them, and relations between them. How much and what kind of knowledge is needed depends on the tasks these knowledge structures are intended to support.

In the Information Science community, an *ontology* is generally defined as a formal representation of a shared conceptualization, and so any sufficiently well-defined knowledge structure over which a consensus exists can be seen as an ontology. In that light, a *taxonomy*, whether a biological taxonomy of the animal kingdom or a genre classification of books, is an ontology that captures a strict hierarchy of classes into which individuals can be uniquely classified.

In practice, those who create knowledge structures do not generally call them ontologies unless they encode certain particular kinds of knowledge. For example, ontologies normally differentiate between concepts and their instances. In this survey, we distinguish the three categories of knowledge structure shown in Table 1 according to the kind of information that they encode: term lists, term hierarchies, and semantic databases. In practice, these categories form a loose spectrum: the distinctions are not hard and fast.

Term lists include most dictionaries, vocabularies, terminology lists, glossaries, and lexicons. They represent collections of terms, and may include definitions and perhaps information about synonymy, but they lack a clear internal structure. The various names in the above list imply certain characteristics. For example, 'dictionary' implies a comprehensive, ideally exhaustive, list of words with all possible definitions of each, whereas 'glossary' implies a (nonexhaustive) list of words with a definition of each in a particular domain, compiled for a particular purpose.

Term hierarchies specify generic semantic relations, typically *has-broader* or *has-related*, in addition to synonymy. In this category, we include structures such as thesauri, controlled vocabularies, subject headings, term hierarchies, and data taxonomies. The word 'taxonomy' implies a structure defined for the purposes of classification in a particular domain (originally organisms), whereas 'thesaurus' implies a comprehensive, ideally exhaustive, listing of words in groups that indicate synonyms and related concepts. However, in many circumstances the names are used interchangeably. According to standard definitions of taxonomy and thesaurus, antonym (opposite meanings) is not required information in either, nor is it supported by common formats. However, it is included in many traditional thesauri—notably Roget's. Subject headings are hierarchical structures that were originally developed for organizing library assets; their structure closely resembles taxonomies and thesauri. Most encyclopedias are best described as glossaries with immense depth and coverage. Wikipedia, however, can be viewed as a taxonomy, because its articles are grouped hierarchically into categories and their definitions include hyperlinks to other articles that indicate generic semantic relationships.

Semantic databases are the most extensive knowledge structures: they encode domain-specific knowledge, or general world knowledge, comprehensively and in considerable depth. Besides differentiating between concepts and their instances, a typical ontology falling into this category would also encode specific semantic relations, facts and axioms.

TABLE 1 | Three Categories of Knowledge Structures

	Term Lists	Term Hierarchies	Semantic Databases
What knowledge structures belong here?	Lexicons, glossaries, dictionaries	Taxonomies, thesauri, subject headings	Ontologies, knowledge repositories
What are examples of such structures?	Atis Telecom Glossary	MeSH, LCSH, Agrovoc, IPSV, and many more	CYC, GO, DBpedia YAGO, BabelNet
How are semantic units represented?			
As terms (with optional descriptions)	✓	✓	
As concepts			✓
Which semantic relations are represented?			
Equivalence: synonymy and abbreviations	✓	✓	✓
Antonym		✓	✓
Generic hierarchical relations (*has-broader*)		✓	
Generic associative relations (*has-related*)		✓	
Specific hierarchical relations			✓
Hypernym/hyponym (*is-a*)			
Concepts vs instance (*is-instance-of*)			
Nonhierarchical relations			✓
e.g., Meronymy (*has-part*)			
Specific semantic relations			✓
e.g., *Is-located-in, works-at, acquired-by*			
What additional knowledge is represented?			
Entailment: dog barks *entails* animal barks			✓
Cause: killing *causes* dying			✓
Common sense			✓
What are the example use cases?	Index of specialized terms	Indexing content, exploratory search, browsing	NLP and AI applications
What standards exist for these resources?	–	ANSI/NISO Z39.19, ISO 25964	ISO 24707
What are typical encoding formats?	GlossML (XML)	SKOS (RDF)	OWL, OBO

Many also encode semantic 'common-sense' knowledge, such as disjointness of top-level concepts (*Artifact* vs *Living being*—one cannot be both), attributes of semantic relations like transitivity, and perhaps even logical entailment and causality relations. Although such structures were originally crafted manually and therefore limited in coverage, several vast knowledge repositories, many boasting millions of assertions comprising mainly particular instances and facts, have recently been automatically or semiautomatically harvested from the web.

The more subtle the knowledge to be encoded, the more complex is the task of creating an appropriate knowledge structure. The payback is the enhanced expressiveness that can be achieved when working with such structures, which increases with its complexity. Figure 2 illustrates this relationship in terms of the three categories shown in Table 1 and discussed above.

Figure 2 shows some overlap between the knowledge structures. Of course, this causes confusion: one person might call something a taxonomy, whereas another calls it an ontology. The fact is that some knowledge structures are hard to categorize. The popular lexical database WordNet[1] is unusual in that it describes not only nouns but also adjectives, verbs, and adverbs. It organizes synonymous words into groups (called 'synsets') and defines specific semantic relations between them. Although WordNet was not originally designed as an ontology, recent versions do distinguish between concepts and instances, turning it into a fusion of a lexical knowledge base and what its original creator has referred to as a 'semiontology'.[2] Freebase[3] and DBpedia[4] are knowledge bases in which the vast majority of entries are instances of concepts, defined using specific semantic relations, including temporal and geographical relations and other worldly facts. The Web contains a plethora of domain-specific sources: GeoNames[5] encodes hierarchical and geographical information about cities, regions, and countries; UniProt[6] lists proteins and relates them to scientific concepts such as biological processes and molecular function; there are countless others.

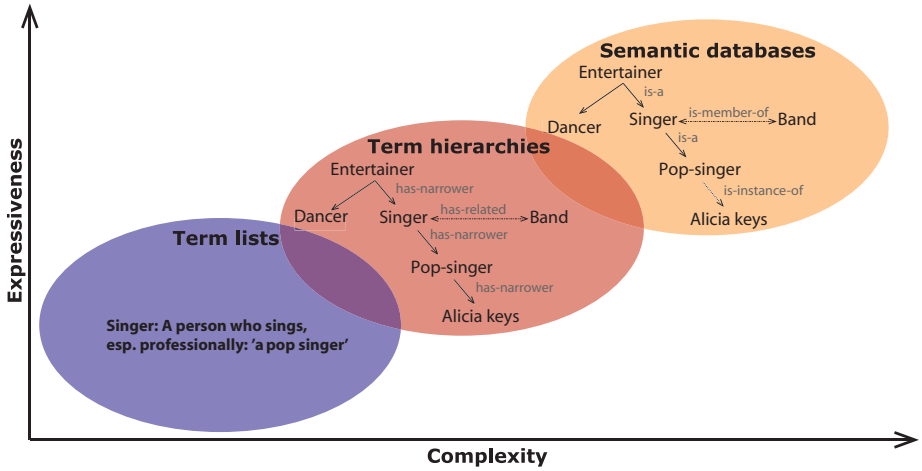

FIGURE 2 | The relation between complexity and expressiveness.

What knowledge structures include is determined by their purpose and intended usage. However, knowledge collected with a particular goal in mind often ends up being redeployed for different purposes. Sources originally intended for human consumption are being re-purposed as knowledge bases for algorithms that analyze human language. WordNet, for example, was created by psychologists to develop an explanation of human language acquisition, but soon became a popular lexical database for supporting natural language processing tasks such as word sense disambiguation, with the ultimate goal of automated language understanding and machine translation. Similarly, Wikipedia,[7] created by humans for humans as the world's largest and most comprehensive encyclopedia, available in many different languages, is being mined to support language processing and information retrieval tasks.

Origins, Standards, and Formats

Endeavors to automate the construction of knowledge structures originate in information retrieval, computational linguistics, and Artificial Intelligence, which all aspire to equip computers with human knowledge. In information retrieval, knowledge is needed to organize and provide access to the ever-growing trove of digitized information; in computational linguistics, it drives the understanding and generation of human language; and in artificial intelligence, it underpins efforts to make computers perform tasks that one would normally assume to require human expertise.

The key problems in information retrieval are determining which terms that appear in a document's text should be stored in the index,[8] and matching terms in users' queries to these terms.[9] Modern terminology extraction techniques still use basic text processing such as stopword removal and statistical term weighting, which originated in the early years.

Early computational linguistics research explored large machine-readable collections of text to study linguistic phenomena such as semantic relations and word senses,[10] and also addressed key issues in text understanding such as the acquisition of a linguistic lexicon.[11] In language generation, lexical knowledge of collocations, i.e., multiword phrases that tend to co-occur in the same context, is necessary to construct cohesive and natural text.[12] Many of the statistical measures developed over the years for automatically acquiring collocations from text[13] are used for extracting lists of terms worth including in a knowledge structure.

Knowledge engineering, a subfield of Artificial Intelligence, addresses the question of how best to encode human knowledge for access by expert systems.[14] Early expert systems[15,16] were designed with a clear separation between the *knowledge base* and *inference engine*. The former was encoded as rudimentary IF-THEN rules; the latter was an algorithm that derived answers from that knowledge base. As the technology matured, the difficulty of capturing the required knowledge from a human expert became apparent, and the focus of research shifted to techniques, tools, and modeling approaches for knowledge extraction and representation. Ontologies became important tools for knowledge engineering: they formulate the domain of discourse that a particular knowledge base covers. Put more concretely, they nail down the terms that can be reasoned about and define relations between them. Current ontology representation languages emerged from early work on frame languages and semantic nets, such as the

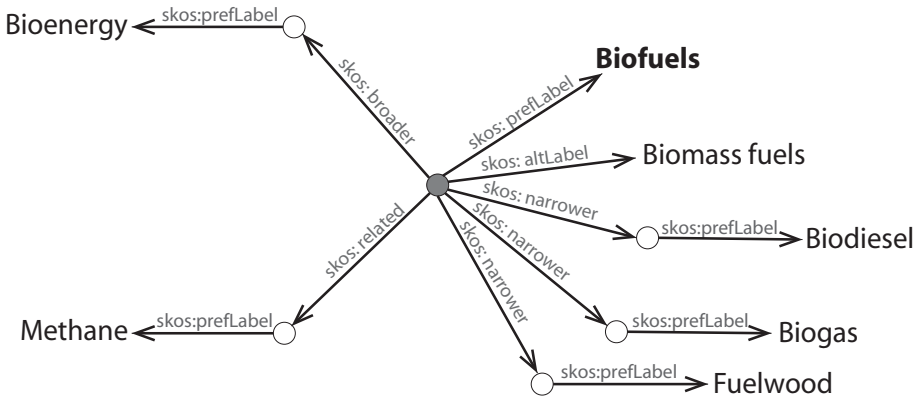

FIGURE 3 | Simple knowledge organization system (SKOS) core vocabulary for the Agrovoc Thesaurus; each circle represents a concept.

KL-One Knowledge Representation System.[17] The notion of Web-enabled ontologies is more recent. Early efforts such as OntoBroker[18] and, in particular, OIL[19] and DAML-ONT,[20] have culminated in the creation of a standardized Web Ontology Language, OWL.[21]

The World Wide Web Consortium (W3C), an international standards organization for the World Wide Web, has endorsed many languages that are used for encoding knowledge structures. Besides OWL, another prominent representation language is the simple knowledge organization system,[22] or SKOS, which is a popular way of encoding taxonomies, thesauri, classification schemes, and subject heading systems in RDF form. Figure 3 shows the SKOS core vocabulary for an example from the Agrovoc Thesaurus[23] vocabulary. Other standards organizations, such as ISO and ANSI, also promote common standards for defining taxonomies and ontologies (see Table 1).

EXISTING TAXONOMIES, ONTOLOGIES, AND OTHER KNOWLEDGE STRUCTURES

There is plethora of knowledge structures, both general and specific. Some have been painstakingly created over the years by groups of experts; others are automatically derived from information on the Web, currently as research projects. The results are freely available or can be obtained for a fee. In some cases there are both free versions and full commercial versions.

Table 2 lists some knowledge sources in various fields, along with the size and year of the latest version. Further examples can be found on the W3C Semantic Web SKOS wiki,[24] by searching the CKAN Data Hub,[25] or by browsing OBO Foundry[26] and Berkeley BOP.[27]

As web standards advance, such structures are becoming increasingly interlinked, gradually expanding the network of 'linked open data'[28] that drives the adoption of the Semantic Web.[29] Figure 4 shows how the definition of *Africa* in the New York Times taxonomy is linked through the *owl:sameAs* predicate to its definition in other sources, such as DBpedia, Freebase, and GeoNames. As well as the enhanced expressiveness that these supplementary definitions bestow, the linkages allow further information to be derived, such as alternative names for Africa in many languages from the GeoNames database.

Historically, those who have created taxonomies and ontologies have not linked them to other knowledge sources. Recently, efforts have been made to rectify this. For instance, the 2012AB release of the unified medical language system (UMLS)[30] integrates 11 million names in 21 languages for 2.8 million concepts from 160 source vocabularies (e.g., GO, OMIM, MeSH, MedDRA, RxNorm, and SNOMED CT), as well as 12 million relations between concepts. Another recent project addressing the same issue is the Bio2RDF.org, which in January 2013 featured 1 billion triples across 19 datasets. Because of the size and complexity of the biomedical domain, rules have been established for integrating inter-related concepts, terms, and relationships. This process is not without errors; new releases appear bi-annually.

In the area of linguistics, most data has been published in proprietary closed formats. A gradual shift is now taking place toward more open linked data formats for representing linguistic data, as proposed, e.g., by Chiarcos et al.[31]

TABLE 2 | Some Publicly Available Knowledge Structures

Name	Field	Built	Size	Year	Source and Year of Latest Version
Term Hierarchies					
LCSH	General	M	337,000 headings	2011	id.loc.gov
MeSH	Biomedical	M	26,850 headings	2013	ncbi.nlm.nih.gov/mesh
Agrovoc	Agriculture	M	40,000 concepts	2012	fao.org/agrovoc
IPSV	General	M	3,000 descriptors	2006	doc.esd.org.uk/IPSV
AOD	Drugs	M	17,600 concepts	2000	etoh.niaaa.nih.gov
NYT	News	M/A	10,4000 concepts	2009	data.nytimes.com
Snomed CT	Healthcare	M	331,000 terms	2012	ihtsdo.org/snomed-ct
Semantic Databases					
WordNet	General	M	118,000 synsets	2006	wordnet.princeton.edu
GeoNames	Geography	M	10,000,000	2012	geonames.org
GO	Bioscience	M	76,000	2012	geneontology.org
PRO	Bioscience	M	35,000	2012	pir.georgetown.edu/pro
Cyc	General	M	500,000 concepts; 15,000 relations; 5,000,000 facts	2013	cyc.com
Freebase	General	M	23,000,000	2013	freebase.com
WikiNet	General	A	3,400,000 concepts; 36,300,000 relations	2010	h-its.org/english/research/nlp
DBpedia	General	A	3,770,000 concepts; 400,000,000 facts	2012	dbpedia.org
YAGO	General	A	10,000,000 concepts; 120,000,000 facts	2012	yago-knowledge.org
BabelNet	General	A	5,500,000 concepts; 51,000,000 relations	2013	lcl.uniroma1.it/babelnet

M stands for manual and A for automated creation.

THE STAGES IN MINING MEANING

Knowledge structures are often constructed to support particular tasks. The application dictates how expressive the representation should be, and what level of analysis is needed. Buitelaar et al.[32] present an 'ontology learning layer cake' which divides the process of ontology learning into separate tasks in ever-increasing complexity as one moves up the hierarchy, with the end product of each task being a more complex knowledge structure. Our own analysis loosely follows this layered approach, reviewing what can be achieved in a way that proceeds from simple to more complex semantic analysis, corresponding roughly to moving upwards and to the right in Figure 2.

Africa

http://data.nytimes.com/66226910773031526401

geo:lat	7.18810087117902
geo:long	21.09375
nyt:associated_article_count	633
nyt:first_use	2004-09-05
nyt:latest_use	2010-06-13
nyt:number_of_variants	3
owl:sameAs	http://data.nytimes.com/africa_geo
owl:sameAs	http://dbpedia.org/resource/Africa
owl:sameAs	http://rdf.freebase.com/ns/en.africa
owl:sameAs	http://sws.geonames.org/6255146/
rdf:type	http://www.w3.org/2004/02/skos/core#Concept
skos:inScheme	http://data.nytimes.com/elements/nytd_geo
skos:prefLabel - en	Africa

FIGURE 4 | Entry for 'Africa' in the New York Times taxonomy.

From Text to Terms

Identifying relevant terminology in a particular domain, possibly defined extensively by a given document collection, is a preliminary step toward constructing more expressive knowledge structures such as taxonomies and ontologies.[33] Riloff and Shepherd[34] argue that it is necessary to focus on a particular domain because it is hard to capture all specific terminology and jargon in a single general knowledge base. One approach to creating a lexicon for a domain like *Weapons* or *Vehicles* (their examples) is to identify a few seed terms (e.g., *bomb*, *jeep*) and iteratively add terms that co-occur in documents.[34] Another is to use statistics, in a similar way to *keyword extraction*, to identify a handful of the most prominent terms in a document.[35] The resulting lists show valuable for tasks like back-of-the-book indexing, where algorithms can potentially eliminate labor-intensive work by professional indexers. A decision which terms are worth including is subjective, of course, and even experts disagree on what should be included in dictionaries or back-of-the-book indexes. Hence, only low accuracy can be achieved—around 50% for terminology extraction[36] and 30% for back-of-the-book indexing.[35]

From Terms to Meaning

Once prominent terms in documents have been identified, the next step is to determine their meaning. By examining a term's context, one can determine its semantic category. *Named entity recognition* is a particular case, where proper nouns that correspond to categories such as People, Organizations, Locations, and Events are determined.[37] Other possible categories include prominent entity types in a given domain: drugs, symptoms and organisms in biomedicine; police, suspects, and judges in law enforcement. Semantic relations between terms and their categories derived in this manner can be used to build new taxonomies or expand existing ones.

Although semantic categories restrict what a given term means, they do not pinpoint its precise denotation. John Smith is a Person, but there are many John Smiths; Frankfurt Police may refer to police stations in different cities. Meanings are what encyclopedias, dictionaries, and taxonomies define, so one way of expressing a term's denotation is to link it to such a source using a unique identifier supplied by that source. For most terms, disambiguation based on context is necessary to determine the correct identifier and discard inappropriate meanings.

A popular trend is to automatically link terms in running text to articles in Wikipedia, a process called *Wikification* or *entity linking*.[38–40] Figure 5 illustrates some of the issues involved in relating a short fragment of text to Wikipedia: ambiguity (shown for just four terms here); overlapping concept references; selection of informative links [many potential links have been omitted from the figure, to concepts such as *Six* (*number*), *Half* (*one half*), *Have* (*property*), *The* (*grammatical article*)]. Such systems exploit the extensive definitions and rich hyperlinking exhibited by Wikipedia articles and achieve around 90% accuracy on Wikipedia articles and 70% on non-Wikipedia text. The likely reason for lower accuracy on the latter is that text often refers to entities that are not included in Wikipedia—e.g., names of ordinary people, rather than celebrities. Recent research has specifically addressed the question of detecting such entities.[40]

In biology, a common task is to identify gene and protein names in text and link them to sources such as Entrez Gene[41] or Uniprot, a process called *gene normalization*. The results from the BioCreative II competition in 2008 show that individual systems typically achieve an accuracy of 80%; however, combining systems using a voting scheme can increase performance to over 90%.[42] DBpedia is another popular resource for annotating words in text with their denotation, a task for which current techniques report around 60% accuracy.[43–45] DBpedia is part of the linked data cloud, so results can be expressed as RDF triples, making them easy to query and re-use in other applications.

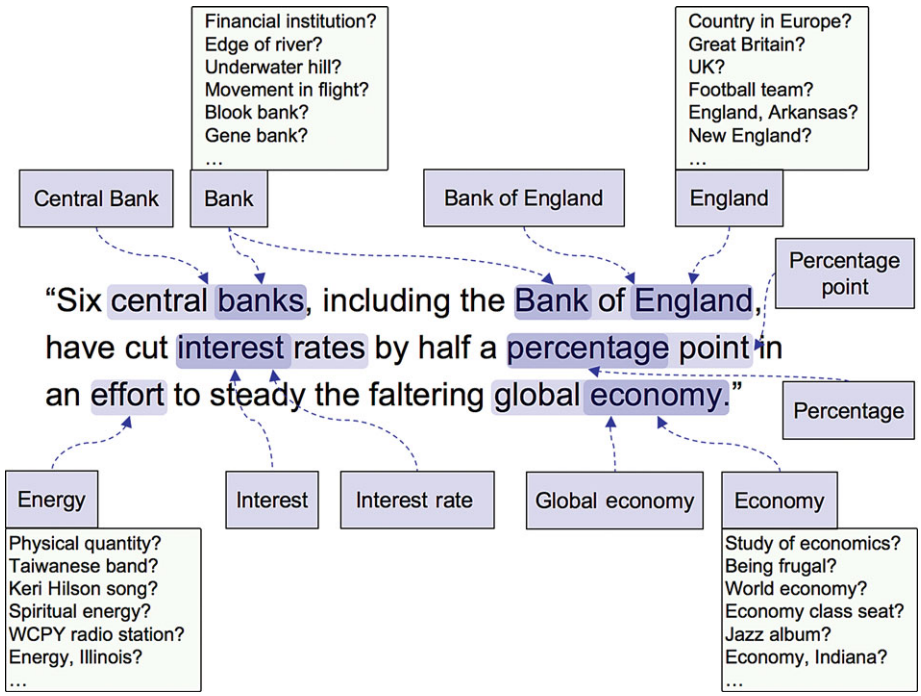

FIGURE 5 | Relating a fragment of text to Wikipedia.

From Terms to Hierarchies

Disambiguating terms in documents is the first step in creating a custom taxonomy or ontology that underpins the knowledge expressed in a particular document collection. Many projects strive to organize extracted terms automatically into hierarchical structures by determining pairs of terms where one has broader meaning than the other.

It is possible that all hierarchical relations extracted in this manner constitute a single connected structure, a *taxonomy*. More likely, the result is a forest of disconnected smaller trees, referred to as *facets*, *faceted taxonomies*, *faceted metadata*, or *dynamic taxonomies*.[46,47] Such structures can facilitate browsing a document collection by successively refining the focus of a search. For example, when seeking blogs that review gadgets, one may choose to narrow an initial search by the type of gadget (e.g., mobile phone), then by manufacturer (e.g., Apple), and finally by model (e.g., iPhone 4s). In such applications, it is necessary to build an index that records which terms appear in which documents. When creating facets, some broader terms are given preference over others because they seem to be more informative when navigating search results. Ideally, the facets that are displayed would depend on the query, e.g., a search for *us movies* would result in facets such as actor, director, and genre.

Several techniques of linguistic analysis can help identify hierarchical relations between words and phrases: lexico-syntactic patterns, co-occurrence analysis, distributional similarity computation, and dependency parsing. These techniques are reviewed in the next section. When extracting hierarchical relations, the goal may be broader than simply to organize a document collection. Extracted taxonomies are an intermediate step in constructing larger and more expressive knowledge structures, or in enlarging existing ones.[48,49]

Evaluating hierarchies is a difficult task, and quality can rarely be captured by a single metric. Some researchers compare the hierarchy they produce to existing ones in terms of coverage[48] or in terms of its ability to support particular tasks.[50] Others recruit human judges to estimate the quality of a hierarchy, either overall or in terms of particular relations.[46,49]

Relations and Facts Extraction

Other kinds of semantic relation can be extracted from text, not just hierarchical ones. An extensive body of research in *information extraction* and *text mining* strives to automatically detect all the relations listed in Table 1. The ultimate goal is to build a fully comprehensive database of knowledge,[51] preferably one that can be improved upon iteratively. This is an automatic analog of the long-standing Cyc project,[52] which has manually assembled a comprehensive ontology and knowledge base of everyday commonsense knowledge that also evolves over time. Use cases range from answering questions to automatically acquiring new knowledge—e.g., by inferring it from causal relations.

Perhaps the ultimate test of a comprehensive knowledge base is its ability to respond to questions on a wide variety of topics. A striking example of a comprehensive and successful question-answering system is Watson,[53] created by scientists at IBM, which outperformed human contestants to win the Jeopardy quiz show. It combines a variety of content analysis techniques that merge information extracted from the Web with knowledge that is already encoded in resources such as WordNet, DBpedia, and YAGO.[54] Figure 6 illustrates the gradual improvement in its performance from version to version: the last version shown outperformed many people, shown as dots in a 'Winners Cloud'.[55] Another standout example is Wolfram Alpha,[56] an impressive system to which people can pose factual questions or calculations. However, in this case the answers already reside in various databases in structured form: the challenge is not to extract facts from text but rather to translate natural language questions into conventional database queries.

In the biomedical domain, relations extracted from diverse sources are mined to generate hypotheses that stimulate the acquisition of new knowledge. The field of *literature-based discovery* began in 1986 when a literature review of two disparate fields revealed a connection between Raynaud's syndrome and fish oil, on the basis that the former presents high blood viscosity and the latter is known to reduce blood viscosity.[57] This established the Swanson linking model. Following this seminal work, several groups have worked on automated approaches for literature-based discovery,[58,59] which has now spread beyond the biomedical field into applications such as water purification.[60]

Many groups have extracted relationships from biomedical documents, including protein–protein[61] interactions and interactions both between drugs[62] and between genes and drugs.[63] Recently, with the rise of 'big data' and systems biology approaches, biologists are building vast networks of genes, proteins, and often other entities (chemicals, metabolites). This enables them to investigate biological processes at

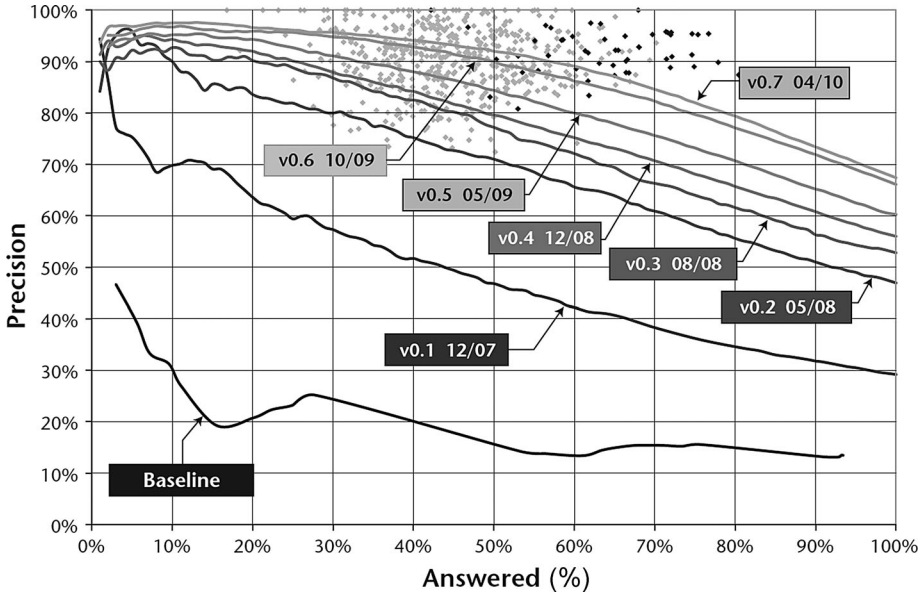

FIGURE 6 | Improvement in Watson's performance (the dot cloud shows people's performance).

the level of functional modules rather than individual proteins.[64]

AUTOMATIC CONSTRUCTION OF KNOWLEDGE STRUCTURES

Approaches to automatically constructing knowledge structures can be grouped by the categories in Table 1. Here, we summarize the techniques used in research projects over the past two decades.

Glossaries, Lexicons, and Other Term Lists

Automatic identification of words and phrases that are worth including in a glossary, lexicon, back-of-the-book index, or simply a list of domain-specific terminology, is a first step in constructing more comprehensive knowledge structures. Here, three main questions of interest are:

1. Which phrases appearing in text might represent terms?
2. When does a phrase become a term?
3. How can a term's meaning in a given context be determined, and synonymous phrases be found?

When detecting terms in text, attention can be restricted to certain words and phrases, excluding others from further consideration. For example, one might ignore phrases such as *list of*, *including* or *phrases that are worth*, and focus only on phrases that could denote terms, e.g., *automatic identification*, *glossary*, and *knowledge structures*. An *n*-gram is a sequence of *n* consecutive words, where *n* ranges from 1 up to a specified maximum. Simply extracting all *n*-grams and discarding ones that begin or end with a stopword yields all valid terms but includes numerous extraneous phrases. Alternatively, one can determine the syntactic role of each word using a part-of-speech tagger and then either seek sequences that match a predetermined set of tag patterns, or identify noun phrases using shallow parsing. This yields a greater proportion of valid terms, but inevitably misses some. Figure 7 compares two sets of candidate phrases, one identified using the *n*-gram extraction approach; the other using shallow parsing. Some systems employ named entity recognition tools to identify noteworthy names. A comprehensive comparison of various methods for detecting candidate terms concluded that

NEJM usually has the highest impact factor of the journals of clinical medicine.

N-grams:	**Noun Phrases:**
NEJM	NEJM
Highest	Highest impact factor
Highest impact factor	Journals
Impact	Clinical medicine
Impact factor	
Journals	
Journals of clinical	
Clinical	
Clinical medicine	
Medicine	

FIGURE 7 | *n*-Grams versus noun phrases.

a combination of *n*-grams and named entities works best.[35]

Having gathered candidate phrases from the text, the next task is to determine whether or not each one is a valid term. Current methods are statistically driven, and can be divided into two categories. The first ranks candidates using criteria such as the *t*-test, C-value, mutual information, log likelihood, or entropy. There are two slightly different classes of measure: lexical cohesion (sometimes called 'unithood' or 'phraseness'), which quantifies the expectation of co-occurrence of words in a phrase (e.g., *back-of-the-book index* is significantly more cohesive than *term name*); and semantic informativeness (sometimes called 'termhood'), which highlights phrases that are representative of a given document or domain. Occurrence counts in a generic corpus are often used as a reference when computing such measures. Some researchers evaluate different ranking methods and select one that best suits their task;[36,65] others combine unithood and termhood measures using a weighted sum[35] or create a new metric that combines both.[66]

The second way of identifying terms uses bootstrapping. First, seed terms for a given semantic category, e.g., *Vehicles* or *Drugs*, are determined, either manually or automatically. Further terms are identified by computing their co-occurrence probability with the seed terms, and the process is repeated iteratively. The idea was proposed by Riloff and Shepherd[34] and has been refined and extended by others over the years.[67–69] This second approach is more semantically focused than the first, being seeded with terms that denote specific semantic categories—whereas the first approach seeks any terms that are generally salient. In one method, seed terms are determined randomly from the pool of content words, which are words that occur within certain frequency thresholds, and clustered into semantic categories using pattern analysis.[70]

Once terms have been identified, their variants, paraphrases, and synonyms must be grouped under the same entry. Bourigault and Jacquemin[71] use extended part-of-speech patterns to determine syntactic variants like *cylindrical bronchial cell* and *cylindrical cell*, or *surface coating* and *coating of surface*. Park et al.[66] divide such variants into five types that can be detected automatically by linguistic normalization tools: (1) symbolic (*audio/visual input* and *audio-visual input*); (2) compounding (*passenger airbag* and *passenger air bag*); (3) inflectional (*rewinds* and *rewinding*); (4) misspelling (*accelarator* and *accelerator*); and (5) abbreviations. Csomai and Mihalcea[35] determine whether two terms are lexical or syntactic

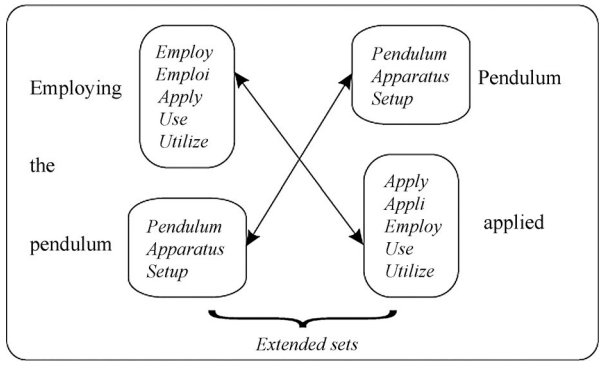

FIGURE 8 | Example of paraphrase identification.

paraphrases by checking for nonempty intersection between a set of labels for each term that comprises the stem and WordNet synonyms of every nonstopword it contains; Figure 8 shows an example for the terms *employing the pendulum* and *pendulum applied*. It would be interesting to study which types of variant are most common in practice, and devise schemes that account for all types.

Taxonomies, Thesauri, and Other Hierarchies

Some work on extracting terminology from text takes account of basic broader/narrower relations between terms. For example, when Riloff and Shepherd[34] bootstrap term extraction for the category *Vehicles*, a two-level taxonomy with a single root and many leaves is formed. Subsequent extraction of terms for related categories (e.g., *Vehicle parts*) could add other branches, and so on, iteratively.

The research surveyed below focuses on generating taxonomies rather than lists of terms, the goal being either to deduce a multilevel hierarchical structure for use when browsing documents and suggesting search refinements, or as an intermediate step when constructing more complex structures. We identify two strands of work: creating taxonomies from plain text and carving hierarchies from existing knowledge structures.

Taxonomic relations can be derived from text using a pattern-based approach. In seminal early work, Hearst[72] mined Grolier's encyclopedia using a handful of carefully chosen lexico-syntactic patterns, shown in Table 3. According to human judges, 52% of the relations extracted were 'pretty good'—but the technique was only 28% accurate on a different corpus (Lord of the Rings). Many researchers have extended this work. For example, Cederberg and Widdows[73] use Latent Semantic Analysis to

TABLE 3 | Lexico-Syntactic Patterns for Extracting Relations from Text

Pattern	Matching Text	Extracted Relation
NP_0 such as {NP_1, NP_2 . . . , (and|or)} NP_n	. . . found in fruit, such as apple, pear, or peach, . . .	apple is-a fruit; pear is-a fruit; peach is-a fruit
such NP as {NP,}* {(or|and)} NP	. . . works by such authors as Herrick, Goldsmith, and Shakespeare . . .	Herrick is-a author; Goldsmith is-a author; Shakespear is-a author
NP {, NP}* {,} or other NP	. . . bruises, wounds, broken bones, or other injuries . . .	bruise is-a injury; wound is-a injury; broken bone is-a injury
NP {, NP}* {,} and other NP	. . . temples, treasuries, and other civic buildings . . .	temple is-a civic building; treasury is-a civic building
NP {,} including {NP,}* {or|and} NP	. . . countries, including Canada and England . . .	Canada is-a country; England is-a country
NP {,} especially {NP,}* {or|and} NP	. . . most European countries, especially France, England, and Spain . . .	France is-a European country; England is-a European country; Spain is-a European country

compute the similarity between hyponym pairs, reducing the error rate by 30% by filtering out dissimilar and therefore incorrect pairs. They observed that Hearst's patterns that indicate hyponymy may also have other purposes. For example, *X including Y* may indicate hyponymy (e.g., *illnesses including eye infections*) or membership (e.g., *families including young children*) depending on the context. They also noticed that anaphora can block the extraction of broader terms that appear in a preceding sentence (e.g., 'A kit such as X, Y, Z will be a good starting kit', where the previous sentence mentions *beer-brewing kit*). Snow et al.[74] replaced Hearst's manually defined patterns by automatically extracted ones, which they generalized. The input text was processed by a dependency parser, and dependency paths were extracted from the parse tree as potential patterns, the best of which were selected using a training set of known hypernyms. These patterns were reported to be more than twice as effective at identifying unseen hypernym pairs as those defined by Hearst. Interestingly, this technique can supply quantitative evidence for manually crafted patterns: e.g., it shows that *X such as Y* is a significantly more powerful pattern than *X and other Y*. Cimiano et al.[75] also use lexical knowledge, but instead of searching for patterns they apply dependency parsing to identify attributes. For example, *hotel, apartment, excursion, car*, and *bike* all have a common attribute *bookable*, whereas *car* and *bike* are *drivable*. A 'formal concept analysis' technique is then used to group these terms into a taxonomy based on these attributes.

Other approaches use statistics rather than patterns to identify hierarchies in text. Pereira et al.[76] perform a distributional analysis of the words that appear in the context of a given noun, and group them recursively using a clustering technique. Cluster labels are determined from a centroid analysis. Inspired by the cosine similarity metric in information retrieval, Caraballo[77] created vectors from words that co-occur within appositives and conjunctions of a given pair of nouns in parsed text. They built a taxonomy bottom-up by connecting each pair of most similar nouns with place-holder parent node and then labeling these place-holder nodes with potential hypernyms derived using Hearst's patterns. The labels can be sequences of possible hypernyms, e.g., *firm/investor/analyst*. The final step is to compress the tree into a taxonomy. Sanderson and Croft[78] use subsumption to group terms into a hierarchy. If one term always appears in the same document as another, and also appears in other documents, they assume that the first term subsumes the second, i.e., it is more generic. About 72% of terms identified in this way were genuine hierarchical relations. Yang and Callan[79] compare various metrics for taxonomy induction by implementing patterns, co-occurrences, contextual, syntactic, and other features commonly used to construct a taxonomy, and evaluating their effectiveness on WordNet and Open Directory trees. They conclude that simple co-occurrence statistics are as effective as lexico-syntactic patterns for determining taxonomic relations, and that contextual and syntactic features work well for sibling relationships but less so for *is-a* and *part-of* relations.

When text becomes insufficient, researchers turn to search engines. Velardi et al.[80] focus on lexical patterns that indicate a definition (*X is a Y*), but as well as matching sentences in the original corpus they also collect definitions from the Goole query *define: X* and online glossaries. Kozareva and Hovy[81] suggest constructing search queries using such lexico-syntactic patterns and then analyzing web search engine results

to find broader term for a given term. A similar approach is also used in the extractor module of Etzioni et al.'s[51] ontology KnowItAll (see next section).

Creating Hierarchies Using Relations in Existing Sources

An alternative approach is to use existing knowledge resources such as WordNet or Wikipedia to drive the extraction of a taxonomy, with or without a document collection in mind. Goals range from adding new concepts and relations to existing structures to inducing custom hierarchies from large and comprehensive resources. Note that in this case, the resulting hierarchy contains concepts rather than terms, because the original sources encode these concepts.

Vossen[82] describes how WordNet can be augmented with technical terms. In a corpus of technical writing, he identified noun phrases whose head noun (or noun phrase) matches an existing WordNet entry, and grouped them by common ending. For example, he extended *Technology* to *Printing technology*, and again to *Inkjet printing technology*. He showed that parts of WordNet can be trimmed before this extension to reduce ambiguity, and recommended trimming the upper WordNet classes too. Snow et al.[48] also extended WordNet, but without focusing on any particular domain. Using their earlier method,[74] they harvested many hypernym pairs missing from WordNet, and proposed a probabilistic technique that added 10,000 such pairs with an accuracy of 84%.

Stoica et al.[46] induce a taxonomy from WordNet, focusing on terms mentioned in a given document collection—they used a set of recipes—to support faceted query refinement. They reduced WordNet's hierarchy to a specialized structure intended to support this particular document collection, as illustrated in Figure 9, and their experimental subjects judged it to be significantly more useful than trees generated by Sanderson and Croft's[78] subsumption technique. In the domain of news, Dakka and Ipeirotis[47] noticed that typical facet categories rarely appear in news articles. They used a named entity extraction algorithm in conjunction with WordNet and Wikipedia as a source of terms, which they extended with frequently co-occurring context terms from other resources. They then identified context terms that are particularly common in the news, and constructed a final taxonomy using the subsumption technique.[78] Medelyan *et al.*[83] also describe a method for creating taxonomies for specific document collections. They suggest carving a focused new taxonomy from as many sources as possible: Wikipedia, DBpedia, Freebase, and any number of existing taxonomies in the domain of interest. Heuristics that take account of term occurrences across different documents help select relevant hierarchical relations from the many that are available.

Others extract generic or custom taxonomies from Wikipedia. Observing that its category network includes relations of many types, from strong *is-a* (*Capitals in Asia* and *Capitals*) to weak associations (*Philosophy* and *Beliefs*), Strube and Ponzetto[50] induced a taxonomy by automatically categorizing them into *is-a* and *not-is-a*. They achieved 88% accuracy by combining pattern-based methods with category name analysis.

Ponzetto and Navigli[84] noticed that Wikipedia's category structure copes particularly badly with general concepts. For example, *Countries* is categorized under *Places*, which is in turn categorized under *Nature*: this makes subsumption nontransitive. As a solution they propose to merge the top levels of WordNet with the lower levels of the Wikipedia category structure. The next section describes other attempts to merge various sources into new and complex knowledge structures.

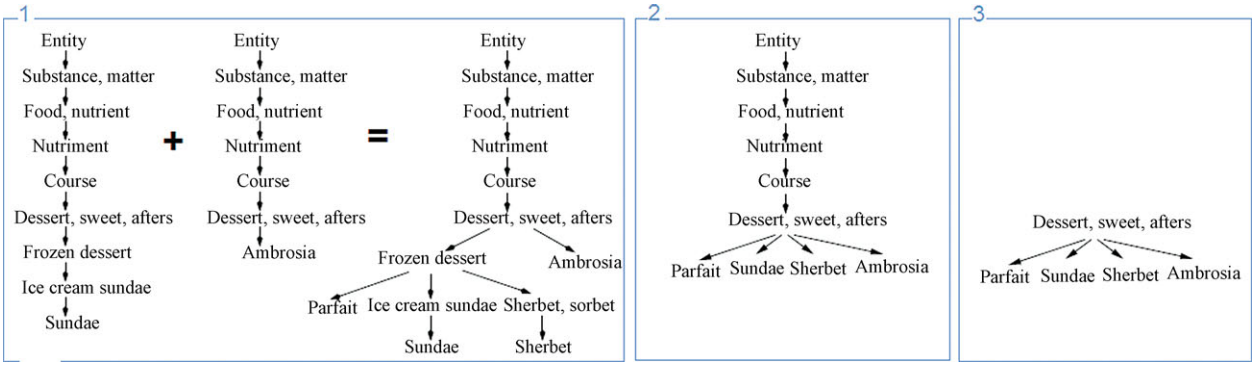

FIGURE 9 | (1) Merging, (2) compressing, and (3) pruning upper levels of WordNet's hypernym paths into a facet hierarchy.

A detailed overview of approaches to generating knowledge structures from collaboratively-built semistructured sources like Wikipedia is provided by Hovy et al.[85] They argue that Wikipedia is particularly suited to this task, not only because of its size and coverage, but also because it is current and covers many languages. They also briefly mention research on inducing ontologies from Wikipedia, which we cover in the next section.

Ontologies, Knowledge Repositories, and Other Semantic Databases

Constructing ontologies is a massive, labor-intensive, and expensive undertaking. Since the early 1990s, various supporting methodologies have been devised. For example, CommonKADS,[86] the European de facto standard for knowledge analysis and knowledge-intensive system development, covers all aspects of ontology development, management, and support, and many major companies in Europe, the United States and Japan have either adopted it in its entirety or partly incorporated it into existing methods. However, though methods such as CommonKADS are very powerful and often come with tool support to assist the ontology engineer, they are, in essence, manual technologies: they still require a knowledge engineer to put in significant amounts of work to shape the ontology.

Consequently, many researchers have turned to automating these processes. Some have developed a variety of methods that combine machine learning tools, NLP techniques, and structured knowledge engineering to construct ontologies from text or other sources. Others focus on building tools and workflows in a multidisciplinary approach to ontology creation. There are also initial attempts to learn deep ontological knowledge, such as disjointness between concepts.

Mining Ontologies from Text

Imagine an algorithm that can read large amounts of text and construct an ontology from the information therein, just as people read books to acquire knowledge. It would have to first identify concepts of interest and then learn facts and relations connecting them.

Lee et al.[87] describe a bottom-up approach for learning an ontology from unstructured text. They identify concepts by detecting terms of interest and clustering them based on similarity. Next they use the notion of *episodes* to cluster co-occurring concepts into meaningful events, which they use as a basis for deeper relation extraction. Their approach addresses some unique challenges posed by Chinese language processing.

Poon and Domingos[88] identify concepts and the relations between them in a unified approach. They use a semantic dependency parser to analyze the sentences and then build a probabilistic ontology (rather than a deterministic one) from logical forms of sentences obtained from this parser.

Others extract facts from the Web, although the resulting structures are not necessarily called ontologies: they operate at term rather than at concept level. The University of Washington's KnowItAll,[51] Carlson et al.'s never ending language learning (NELL) project,[89] and Pasca,[90] all utilize masses of unstructured text crawled from the Web to bootstrap the extraction of millions of facts, and report ever-improving quality. KnowItAll extends Hearst's work[72] by connecting individual lexico-syntactic patterns to classes. For example, *NP1 plays for NP2* is a pattern for collecting facts such as instances of the classes *Athlete* and *SportsTeam*, as well as which athletes play for which teams. A probabilistic engine filters the extracted facts based on co-occurrence statistics derived by querying the web. KnowItAll was soon succeeded by TextRunner[91] and ReVerb.[92] Text Runner implemented a domain-independent approach to fact extraction by removing the need to specify patterns manually, instead deriving them automatically from parse trees. ReVerb extracted more accurate relations by identifying verbs and their closest noun phrases in a sentence as candidate facts, and then using a supervised approach for validating these facts. An interesting aspect of NELL is that it runs continuously, and attempts to improve its extraction capabilities every day by learning from what has been extracted previously. It exploits redundancy that comes from different views of the data, and implements a coupled learning technique that simultaneously learns several facts, connected via their arguments.

Constructing Ontologies from Other Sources

As well as text, pre-existing structured sources have been exploited for automatically constructing ontologies. New ontologies can be created by refining the relations defined in an existing source, extending its coverage, or merging multiple sources into one. The most popular sources are Wikipedia[93–95] and WordNet,[82,96,97] although some researchers have also explored the use of glossaries,[98] existing taxonomies and ontologies,[49] and other linguistic resources.[99]

Two automatically constructed knowledge structures, DBpedia[93] and YAGO,[96] extract concepts

and facts from structured and semistructured parts of Wikipedia. The former focuses on Wikipedia's category structure, infoboxes, images, and links. It represents each category as a class, and uses the key-value pairs available in infoboxes as the basis for properties and relations between objects. Because its focus is on a close mapping between the live Wikipedia and a structured representation of that data, it makes little effort to clean up the structure. The latter is somewhat similar—it also uses the structured content of Wikipedia to construct an ontology—but combines this with information extracted from WordNet, using several heuristics to come up with a higher-quality ontological structure. In contrast to DBpedia, it focuses less on accurately reflecting the contents of Wikipedia, and more on synthesizing a high-quality ontological structure that stands on its own. Recently, a new version of YAGO has been released,[100] which also accounts for temporal and spatial information associated with entities. This system is able to support a system that answers questions such as 'Give me all songs that Leonard Cohen wrote after Suzanne' or 'Name navigable Afghan rivers whose length is greater than one thousand kilometers'.

Nastase and Strube[94] use Wikipedia as a source of semantic relations to extend Strube and Ponzetto's work on taxonomy induction by analyzing category names as well as the category structure. Category names often contain references to other Wikipedia articles, and thousands of specific relations can be extracted using carefully crafted patterns—e.g., from the category *Movies Directed by Woody Allen* one can infer that *Annie Hall* is a *Movie* and '*is directed by*' '*Woody Allen*'. Finally, they also harvest associative relations between concepts that are linked in the same sentence of a Wikipedia article description.

Another strand of work is to extend existing structures with ontological relations. Ruiz-Casado[101] mined Wikipedia for new relations to add to WordNet by creating mappings between WordNet synsets and Wikipedia articles, identifying patterns that appear between Wikipedia articles that are related according to WordNet, and using those patterns to find new relations of different types. Sarjant et al.[49] mine Wikipedia for new concepts to add to the Cyc ontology.[52] They argue that Wikipedia's extensive coverage of named entities and domain-specific terminology complements the general knowledge that Cyc contains. Having created mappings between a Cyc concept and a Wikipedia article, they identify other children of that article's category and filter out non-isa relations by checking the article's first sentence and infobox. They added 35,000 specific concepts to Cyc, such as various dog races and the names of well-known personages.

Another interesting application is to multilingual ontology construction. de Melo and Weikum[102] note the value of Wikipedia's interwiki links as a source of cross-lingual information. Unfortunately many of them are incorrect, so they apply graph repair operations to remove incorrect edges based on several criteria. This work led to MENTA,[103] a multilingual ontology of entities and their classes built from WordNet and Wikipedia that covers more than 200 different languages. MENTA uses a set of heuristics for linking connected Wikipedia articles, categories, infoboxes, and WordNet synsets from multiple languages. The resulting weighted links between entities are aggregated in a Markov chain, in a similar manner to the PageRank algorithm. BabelNet[97] is a multilingual lexicalized semantic network and ontology that covers six European languages (Catalan, French, German, English, Italian, and Spanish) and contains 5.5 million concepts and 26 million word senses. Like MENTA, it was created by integrating Wikipedia with WordNet. Instead of analyzing and correcting existing translation links between different Wikipedia versions, they performed automatic mapping by filling in lexical gaps in resource-poor languages with the aid of statistical machine translation. The resulting semantic network includes 365,000 relation edges from WordNet and 70 million edges of underspecified relatedness from Wikipedia. Like WordNet, BabelNet groups words in different languages into synsets, each containing on average 8.6 synonyms.

Workflows and Frameworks for Building Ontologies

Recently, focus has shifted to a multidisciplinary approach to building tools and workflows for ontology creation. Given the unstructured nature of many information sources, particularly the Web, a combination of machine learning tools, NLP techniques, and structured knowledge engineering seems a promising way to support quicker and easier creation of ontologies.

Maedche and Staab[104] introduce a semiautomatic learning approach that combines technologies from classical knowledge engineering, machine learning and NLP into a workflow and toolset that help knowledge engineers to quickly integrate and reuse existing knowledge to build a new ontology. The method encompasses ontology import, extraction, pruning, refining, and evaluation.

TextToOnto[105] is an ontology framework that integrates existing NLP tools like Gate, and implements additional learning algorithms for concept

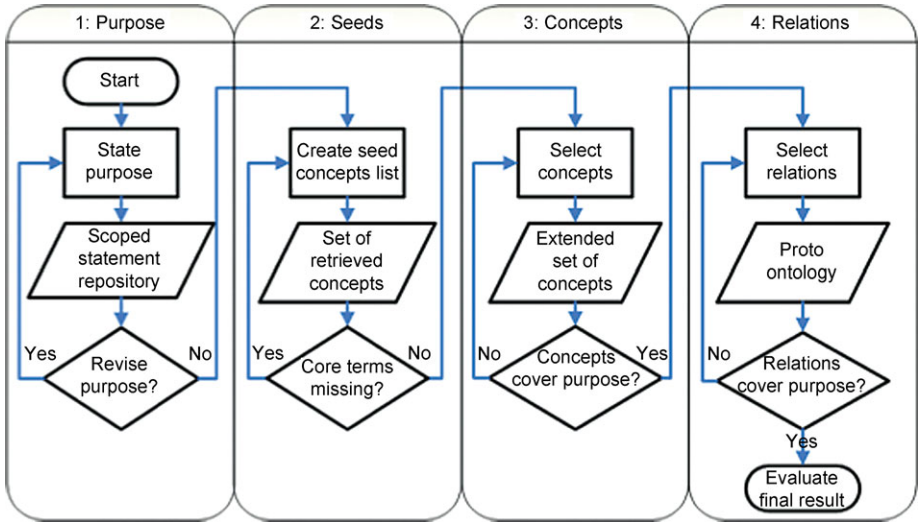

FIGURE 10 | Ontology building workflow in rapid ontology construction (ROC).

relevance and instance-to-concept assignment. All assertions generated are expressed in an intermediate model, which can be visualized or exported into an ontological language of choice.

Koenderink et al.[106] describe rapid ontology construction (ROC), a methodology that distinguishes different stakeholders in the ontology creation process and identifies a workflow for ontology construction based on them. Figure 10 shows an example. ROC includes tools that help automate various steps of the construction process, including selecting likely sources for relevant concepts and using them later to suggest further concepts and relations that should be added.

Gurevych et al.[99] model lexical–semantic information that exists in many different knowledge structures. Their solution unifies all this information into a single standardized model, which also takes into account multilinguality.

Beyond Light-Weight Ontologies

One must note that the result of an automated solution is not always a fully fledged ontology according to the definition in Table 1. Often, so-called 'light-weight ontologies' are constructed that detect classes, instances (or simply concepts), specific semantic relations, and facts. Few approaches are known for automatically detecting common sense knowledge such as disjointness, to be added into a taxonomy. An exception is the work by Völker et al.[107] who learn disjointness from various sources. For example, one of the assumptions made is that if two labels are frequently used for the same item, they are likely to be disjoint, because people tend to avoid redundant labeling. They found that judging disjointness is difficult even for experts, but a supervised system can achieve competitive results.

EVALUATING REPRESENTATIONAL ACCURACY

Evaluating knowledge structures is a crucial step in their creation, and several iterations of refinement are usually needed before finalizing the content and structure. How to evaluate the knowledge structures themselves is still a matter of debate.[108,109] Possible approaches are to compare them with other structures, assess internal consistency, evaluate task based performance, or judge whether they are accurate representations of reality.[110,111]

Most commonly, knowledge structures are evaluated in terms of the *accuracy* of detected concepts, instances, relations, and facts. One begins by comparing automatically determined structures with existing manually produced resources, or by having human judges assess the quality of each item. Then accuracy values are computed using the standard statistical measures used in information retrieval: Precision, Recall, and F-measure. Throughout this paper we quote F-measure values reported by authors as the 'accuracy' of their approach, because these reflect in a single number both aspects of performance, namely how many of the automatically identified items are correct, and how many of the correct items are found.

Another popular way of evaluating knowledge structures is through *task-based performance*, which tests their usability for particular tasks. This includes ease of use, time taken, expertise required, and results achieved. For example, Dakka and Ipeirotis[48]

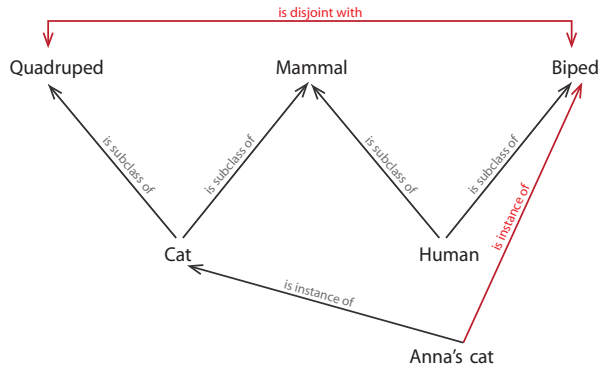

FIGURE 11 | Example of logical inconsistency in an ontology.

designed a study in which users are asked to locate a news item of interest using automatically generated facet hierarchies. The authors report user satisfaction after using the hierarchy, and observations on their interaction with the system. Strube and Ponzetto[50] address the task of computing semantic similarity and compare the accuracy of standard metrics, whether they were relying on WordNet or their automatically generated WikiRelate taxonomy.

Internal consistency is particularly important in ontology learning. For example, logical consistency validates whether an ontology contains contradictory information. Figure 11 shows how adding a new assertion into an ontology (*ANNA'S CAT* is instance of *BIPED*) results in an inconsistency, because *BIPEDS* (walks on two legs) and *QUADRUPED* (walks on four legs) are disjoint.[107] Consistency can be assessed using a variety of metrics, such as clarity, coherence, competency, consistency, completeness, conciseness, expandability, extendibility, minimal ontological commitment, minimal encoding bias, and sensitiveness.

Other, less application-oriented, metrics, have been discussed in the literature. For example, when *comparing knowledge structures* one could analyze structural resemblance. Structure can be compared by measuring the distance between two concepts, represented as nodes in the ontology graph structure, based on shortest path (parsimony), common ancestors, and offsprings, and the degree of branching. For example, Maynard et al.[112] argue that the longer a particular taxonomic path from root to leaf, the more difficult it is to achieve annotation consistency, e.g., indexing consistency. Metrics have also been devised for measuring the breadth and depth of ontologies for the purpose of comparing them with one another within a specific discourse or domain.

Representation of reality (in practice, a subset of reality defined by a particular domain or document collection) is another possible evaluation parameter. It can be judged by measuring the usage frequency of real-world concepts, the alignment of concepts to real-world entities, or by comparing the rate of change in the knowledge structure with that of the real world in terms of the number of concepts added, deleted, or edited.[113] Such evaluation is subjective and can only be accomplished by domain experts.

RESEARCH TRENDS

Proceeding from simple to more comprehensive knowledge structures, here are some salient trends. For term lists, no single method for term detection stands out, and studies comparing the results of different methods conclude that the best choice depends on the overall task goal.[13,36] Early approaches that use hand selected seed terms are being superseded by ones that adopt machine learning techniques to determine an appropriate set of seed terms. For inferring hierarchies from text, researchers still apply patterns to text, but have abandoned manually selected patterns in favor of ones derived automatically via methods such as dependency parsing. This, in conjunction with learning the most effective patterns from data, has doubled accuracy compared with manual selection. Surprisingly, statistical co-occurrence has been found to be just as effective as pattern-based methods (Yang and Callan[79]). When inferring hierarchies from other sources, there is a clear trend toward combining the best of both worlds: e.g., deriving upper levels from WordNet or Cyc and lower levels from Wikipedia and Linked Data sources. Moreover, the focus of research seems to be shifting from generic taxonomies toward the creation of custom structures suitable for browsing specific collections.

Several trends specific to ontologies can be discerned. Whether learning from text or from the web, the challenge is to devise effective pattern extraction methods. More refined methods are suggested each year. When extracting ontologies from existing sources, bigger seems to be regarded as better. In contrast to the taxonomy research mentioned above, ontological sources compete in terms of size and number of facts extracted. One trend is to combine as much information as possible without losing anything (e.g., WordNet senses, facts, hierarchies, links to original sources, Linked Data). Another is to exploit multilingual information sources and link them into a single huge source (e.g., multilingual WordNet and Wikipedia).

Overall, there is a strong trend toward data-driven techniques that use machine learning to derive the optimal parameters, settings, seed words,

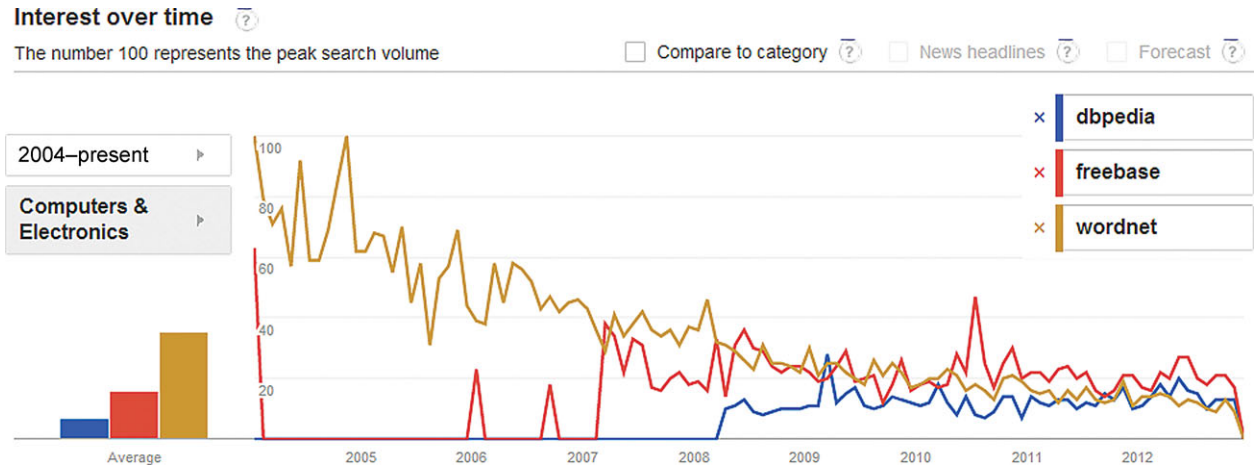

FIGURE 12 | Interest in DBpedia, Freebase, and WordNet.

patterns, etc. The invention of new technologies in machine learning spurs further advances in mining text and other sources for knowledge, which in turn give new insights into the use of human language. Dependency parsing is applied in many different contexts, such as deriving patterns automatically from text, learning common attributes that create hierarchies, and ontology learning. At a practical level there is great interest in formats, frameworks, and APIs that help people work with data, share it with others, support connectivity between sources, and enable it to be easily extendable with new components and knowledge. In practice, researchers tend to re-purpose manually created structures and augment them into larger, more expressive or more specialized resources. Many successful systems combine several sources into one.

Open problems at today's research frontier involve sophisticated ontologies that can work with spatial, temporal, and common sense knowledge. Researchers seem to be leaving behind the inference of entities, facts, simple concepts, and so on, perhaps because these problems are essentially already solved. Instead they are turning attention to the creation of systems (like NELL) that constantly mine the web and continually improve their ability to learn and acquire facts and other knowledge. The robustness of such systems and their sustainability over time are likely to present considerable challenges.

When new, comprehensive sources emerge, researchers gradually abandon others. Figure 12 illustrates how Wikipedia and Freebase have steadily approached and overtaken WordNet as the subject of web searches in the technical field. Another interesting trend can be observed by comparing the number of papers published over time on topics related to the construction of lexicons, taxonomies and ontologies. Figure 13 shows counts from Google Scholar of publications mentioning 'lexicon learning', 'lexicon induction', 'lexicon construction', 'extracting taxonomy', or 'automatically created taxonomy',

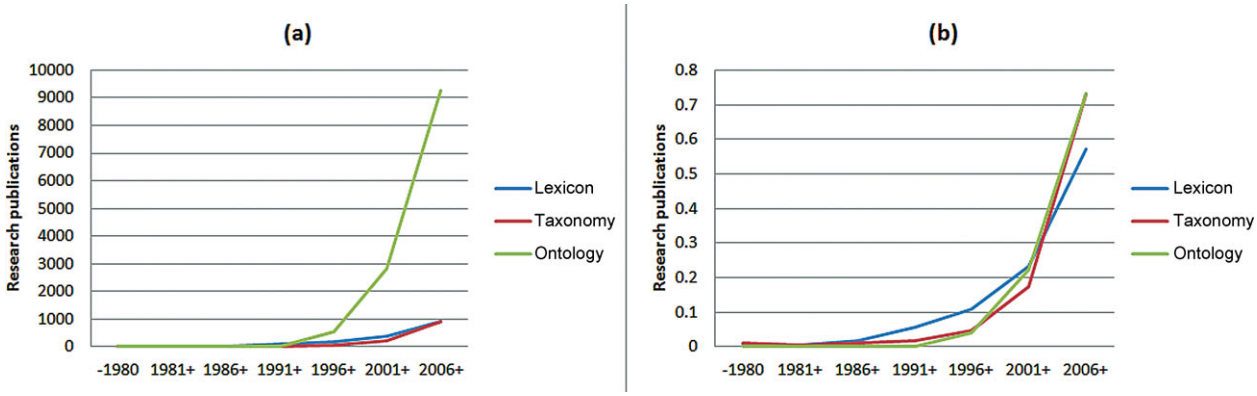

FIGURE 13 | (a) Overall and (b) relative numbers of research publications in recent years.

and corresponding results when *lexicon* is replaced by *taxonomy* and *ontology*, plotted in 5-year intervals from before 1980 to the present day. Automatic construction of ontologies has become significantly more popular, with thousands of publications rather than hundreds. Figure 13(b) compares the relative growth of the three fields, and shows how interest in the construction of lexicons, popular in the 90s, has decayed since 2000 in favor of taxonomy and ontology construction.

CONCLUSION

Over the past decades, researchers have sought the holy grail of a perfect knowledge structure, whether built manually or in some automated fashion. Such a structure would encompass linguistic knowledge of words, phrases, concepts, and their relations; common sense knowledge about how these concepts interact; and factual knowledge that transcends that of the most erudite scholar—although the boundaries between these different types are blurry. Both the complexity and expressiveness of a knowledge structure increase with the amount, variety, and depth of the knowledge it encodes.

Efforts to mine knowledge from text and other sources originated in various fields: information retrieval, as people began to understand the importance of managing digitized data; computational linguistics, as algorithms began to unlock the computer's ability to understand human language; and artificial intelligence, as early expert systems were created to emulate human performance. As time passed, knowledge engineering matured and resulted in new standards and encoding languages, which gradually became widely deployed. Today there are thousands of commercially and publically available lexicons, glossaries, taxonomies, ontologies, and repositories of facts, created both manually and automatically. Many are provided in common formats, with links to one another, or via easily accessible web services. Over the coming years, the rising popularity of the Semantic Web and Linked Data will spur further developments in the linkage and accessibility of existing knowledge structures, which will support ever more powerful applications.

There are numerous reasons for constructing lexicons, taxonomies, ontologies, and other structures. Some researchers attempt to accurately represent the entirety of lexical knowledge and knowledge of language; others focus on constructing a specialized resource for navigating a document collection in a narrow domain; still others set out to collect millions of facts and assertions with the ultimate aim of building a comprehensive oracle or question-answering system.

As automatically constructed knowledge structures become more accurate, comprehensive, and expressive, and with recent attempts to learn even common sense ontological knowledge, we predict the emergence of ever more powerful systems that connect information residing in a variety of sources into a single knowledge base that drives a powerful inference engine. At the same time, the information in many knowledge structures is already available via web services, which frees it from the shackles of a single organization and allows it to be curated, maintained, and updated by its original authors. The key becomes how to combine all this information meaningfully. It is interesting to reflect on how much knowledge about what we know—and more particularly about what we don't know—needs to be captured before we can be confident in being able to support a robust reasoning process. In a sense, this is a contemporary version of the classical 'frame problem' in Artificial Intelligence, which still remains tantalizingly out of reach. Is it possible, in principle, to determine the scope of the knowledge required to derive the full answer to a question, or the full consequences of an action?

REFERENCES

1. Miller GA. WordNet: a lexical database for English. *Commun ACM* 1995, 38:39–41.
2. Miller GA, Hristea F. WordNet nouns: classes and instances. *Comput Ling* 2006, 32:1–3.
3. Freebase. Available at: http://www.freebase.com/. (Accessed December 14, 2012).
4. DBpedia. Available at: http://dbpedia.org/. (Accessed December 14, 2012).
5. GeoNames. Available at: http://geonames.org. (Accessed December 14, 2012).
6. UniProt. Available at: http://www.uniprot.org/. (Accessed December 14, 2012).
7. Wikipedia. Available at: http://wikipedia.org. (Accessed December 14, 2012).
8. Salton G, Lesk ME. Computer evaluation of indexing and text processing. *J ACM* 1968, 15:8–36.

9. Spark-Jones K, Tait, JI. Automatic search term variant generation. *J Doc* 1984, 40:50–66.
10. Church KW, Hanks P. Word association norms, mutual information, and lexicography. *Comput Ling* 1990, 16:22–29.
11. Jacobs P, Zernik U. Acquiring lexical knowledge from text: a case study. In: *Proceedings of the Seventh National Conference on Artificial Intelligence*; 1988, 739–744.
12. Smadja FA, McKeown KR. Automatically extracting and representing collocations for language generation. In: *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*; 1990, 252–259.
13. Evert S. Corpora and collocations. In: A. Lüdeling and M. Kytö, eds. *Corpus Linguistics: An International Handbook*, article 58. Berlin: Mouton de Gruyter.
14. Feigenbaum EA, McCorduck P. *The Fifth Generation*. Reading, MA: Addison-Wesley; 1983.
15. Shortliffe EH. *Computer-Based Medical Consultations: MYCIN*. New York: Elsevier; 1976, Vol. 388.
16. Schank RC, Riesbeck CK. *Inside Computer Understanding*. Hillsdale, N.J.: Lawrence Erlbaum; 1981.
17. Brachman RJ, Schmolze JG. An overview of the KL-ONE knowledge representation system. *Cog Sci* 1985, 9:171–216.
18. Decker S, Erdmann M, Fensel D, Studer R. *Ontobroker: Ontology Based Access to Distributed and Semi-Structured Information*. Karlsruhe, Germany: AIFB; 1998.
19. Broekstra J, Klein M, Decker S, Fensel D, Van Harmelen F, Horrocks I. Enabling knowledge representation on the web by extending RDF schema. In: *Proceedings of the International World Wide Web Conference*; 2001.
20. Hendler J, McGuinness DL. The DARPA agent markup language. *IEEE Intell Syst* 2000, 15:67–73.
21. OWL. Available at: http://www.w3.org/TR/owl-features/. (Accessed December 14, 2012).
22. SKOS. Available at: http://www.w3.org/2004/02/skos/intro. (Accessed December 14, 2012).
23. Agrovoc Thesaurus. Available at: http://aims.fao.org/standards/agrovoc. (Accessed December 14, 2012).
24. W3C Semantic Web SKOS wiki. Available at: http://www.w3.org/2001/sw/wiki/SKOS/Datasets. (Accessed December 14, 2012).
25. CKAN Data Hub. Available at: http://datahub.io/dataset?q=format-skos. (Accessed December 14, 2012).
26. OBO Foundry. Available at: http://obofoundry.org/. (Accessed December 14, 2012).
27. Berkley BOP. Available at: http://www.berkeleybop.org/ontologies/. (Accessed December 14, 2012).
28. Bizer C, Heath, T, Berners-Lee, T. Linked data-the story so far. *Inter J Semantic Web Inform Syst* 2009, 4:1–22.
29. Berners-Lee T. Linked Data. Available at: http://www.w3.org/DesignIssues/LinkedData.html. (Accessed February 12, 2013).
30. UMLS. Available at: http://www.nlm.nih.gov/pubs/techbull/nd12/nd12_umls_2012ab_releases.html. (Accessed December 14, 2012).
31. Chiarcos C, McCrae J, Cimiano P, Fellbaum C. Towards open data for linguistics: linguistic linked data. In: A. Oltramari, et al., eds. *New Trends of Research in Ontologies and Lexical Resources, Theory and Applications of Natural Language Processing*. Berlin Heidelberg: Springer-Verlag; 2013.
32. Buitelaar P, Cimiano P, Magnini, B. Ontology learning from text: an overview. In: Buitelaar P, Cimiano P, Magnini, B, eds. *Ontology Learning from Text: Methods, Evaluation and Applications*. Amsterdam, The Netherlands: IOS Press; 2005, 1–10.
33. Gillam L, Tariq M, Ahmad K. Terminology and the construction of ontology. *Terminology* 2005, 11:55–81.
34. Riloff E, Shepherd J. A corpus-based approach for building semantic lexicons. In: *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics; 1997, 117–124.
35. Csomai A, Mihalcea R. Linguistically motivated features for enhanced back-of-the-book indexing. In: *Proceedings of Annual Meeting of the Association for Computational Linguistics and Human Language Technologies*. Association for Computational Linguistics; 2008, 932–940.
36. Pazienza M, Pennacchiotti M, Zanzotto F. Terminology extraction: an analysis of linguistic and statistical approaches. *Knowl Mining* 2005, 255–279.
37. Nadeau D, Sekine S. A survey of named entity recognition and classification. *Ling Investig* 2007, 30:3–26.
38. Mihalcea R, Csomai A. Wikify!: linking documents to encyclopedic knowledge. In: *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*. ACM; 2007, 233–242.
39. Milne D, Witten IH. Learning to link with wikipedia. In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. ACM; 2008, 509–518.
40. Ratinov L, Roth D, Downey D, Anderson M. Local and global algorithms for disambiguation to wikipedia. In: *Proceedings of the Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics; 2011, 1375–1384.

41. Entrez Gene. Available at: http://www.ncbi.nlm.nih.gov/gene. (Accessed December 14, 2012).
42. Morgan A, Lu Z, Wang X, Cohen A, Fluck J, Ruch P, Divoli A, Fundel K, Leaman R, Hakenberg J, et al. Overview of BioCreative II gene normalization. *Genome Biol* 2008, 9(suppl 2), S3.1–S3.19.
43. Mendes PN, Jakob M, García-Silva A, Bizer C. Dbpedia spotlight: shedding light on the web of documents. In: *Proceedings of the International Conference on Semantic Systems*. ACM; 2011, 1–8.
44. Exner P, Nugues P. Entity extraction: from unstructured text to DBpedia RDF triples. In: *Proceedings of the Web of Linked Entities Workshop in Conjuction with the 11th International Semantic Web Conference*. CEUR-WS; 2012, 58–69.
45. Augenstein I, Padó S, Rudolph S. LODifier: generating linked data from unstructured text. *Semantic Web: Res Appl* 2012, 210–224.
46. Stoica E, Hearst MA, Richardson M. Automating creation of hierarchical faceted metadata structures. In: *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*; 2007, 244–251.
47. Dakka W, Ipeirotis PG. Automatic extraction of useful facet hierarchies from text databases. In: *IEEE International Conference on Data Engineering*. IEEE; 2008, 466–475.
48. Snow R, Jurafsky D, Ng AY. Semantic taxonomy induction from heterogenous evidence. In: *Proceedings of the International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics; 2006, 801–808.
49. Sarjant S, Legg C, Robinson M, Medelyan O. All you can eat ontology-building: feeding Wikipedia to Cyc. In: *Proceedings of the International Joint Conference on Web Intelligence and Intelligent Agent Technology*. IEEE Computer Society; 2008, 341–348.
50. Ponzetto SP, Strube M. Taxonomy induction based on a collaboratively built knowledge repository. *Artif Intel* 2011, 175:1737–1756.
51. Etzioni O, Cafarella M, Downey D, Kok S, Popescu AM, Shaked T, Yates A. Web-scale information extraction in KnowItAll:(preliminary results). In: *Proceedings of the 13th International Conference on World Wide Web*. ACM; 2004, 100–110.
52. Lenat DB, Guha RV, Pittman K, Pratt D, Shepherd M. Cyc: toward programs with common sense. *Commun ACM* 1990, 33:30–49.
53. Ferrucci D, Brown E, Chu-Carroll J, Fan J, Gondek D, Kalyanpur AA, Lally A, Murdock JW, Nyberg E, Prager J, et al. Building Watson: an overview of the DeepQA project. *AI Mag* 2010, 31:59–79.
54. YAGO. Available at: http://yago-knowledge.org. (Accessed December 14, 2012).
55. IBM Watson. Available at: http://www.aaai.org/Magazine/Watson/watson.php. (Accessed December 14, 2012).
56. Wolfram Alpha. Available at: http://www.wolframalpha.com/. (Accessed December 14, 2012).
57. Swanson DR. Fish oil, Raynaud's Syndrome, and undiscovered public knowledge. *Perspect Biol Med* 1986, 30:7–18.
58. Srinivasan P, Libbus B. Mining MEDLINE for implicit links between dietary substances and diseases. *Bioinformatics* 2004, 20:I290–I296.
59. Hristovski D, Peterlin B, Mitchell JA, Humphrey SM. Using literature-based discovery to identify disease candidate genes. *Int J Med Inform* 2005, 74:289–298.
60. Kostoffa RN, Solkab JL, Rushenbergc RL, Wyatt JA. Water purification. *Technol Forecast Soc Change* 2008, 75:256–275
61. Blaschke C, Andrade MA, Ouzounis C, Valencia A. Automatic extraction of biological information from scientific text: protein-protein interactions. In: *Proceedings of the International Conference on Intelligent Systems in Molecular Biology*; 1999, 60–67.
62. Rindflesch TC, Tanabe L, Weinstein JN, Hunter L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. In: *Pacific Symposium on Biocomputing*. Pacific; 2000, 517.
63. Percha B, Garten Y, Altman RB. Discovery and explanation of drug–drug interactions via text mining. In: *Pacific Symposium on Biocomputing*; 2012, 410–421.
64. Krallinger, M, Leitner F, Rodriguez-Penagos C, Valencia A. Overview of the protein–protein interaction annotation extraction task of BioCreative II. *Genome Biol* 2008, 9(suppl 2):S4. Epub September 1, 2008.
65. Wermter J, Hahn U. Finding new terminology in very large corpora. In: *Proceedings of the 3rd international conference on Knowledge capture*. ACM; 2005, 137–144.
66. Park Y, Byrd RJ, Boguraev BK. Automatic glossary extraction: beyond terminology identification. In: *Proceedings of the International Conference on Computational Linguistics*. ACL; 2002, 1–7.
67. Roark B, Charniak E. Noun-phrase co-occurrence statistics for semiautomatic semantic lexicon construction. In: *Proceedings of the International Conference on Computational Linguistics*. ACL; 1998,1110–1116.
68. Thelen M, Riloff E. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In: *Proceedings of the Conference on Empirical Methods in NLP*. ACL; 2002, 214–221.

69. McIntosh T, Curran, JR. Reducing semantic drift with bagging and distributional similarity. In: *Proceedings of the Joint Conference of the ACL and the AFNLP*. ACL; 2009, 396–404.
70. Davidov D, Rappoport A. Classification of semantic relationships between nominals using pattern clusters. In: *Proceedings of Annual Meeting of the ACL on Computational Linguistics*. ACL; 2008.
71. Bourigault D, Jacquemin C. Term extraction + term clustering: an integrated platform for computer-aided terminology. In: *Proceedings of the Conference on European Chapter of the Association for Computational Linguistics*. ACL; 1999, 15–22.
72. Hearst MA. Automatic acquisition of hyponyms from large text corpora. In: *Proceedings of the 14th Conference on Computational Linguistics*; 1992, 539–545.
73. Cederberg S, Widdows D. Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In: *Proceedings of the Conference on Natural Language Learning at HLT-NAACL*. ACL; 2003, 111–118.
74. Snow R, Jurafsky D, Ng AY. Learning syntactic patterns for automatic hypernym discovery. *Adv Neur Inform Proces Syst* 2004, 17:1297–1304.
75. Cimiano P, Hotho A, Staab, S. Learning concept hierarchies from text corpora using formal concept analysis. *J Artif Intel Res* 2005, 24:305–339.
76. Pereira F, Tishby N, Lee L. Distributional clustering of English words. In: *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*. ACL; 1993, 183–190.
77. Caraballo SA. Automatic construction of a hypernym-labeled noun hierarchy from text. In: *Proceedings of Annual Meeting of the ACL on Computational Linguistics*. ACL; 1999, 120–126.
78. Sanderson M, Croft B. Deriving concept hierarchies from text. In: *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM; 1999, 206–213.
79. Yang H, Callan J. A metric-based framework for automatic taxonomy induction. In: *Proceedings of the Joint Conference of the ACL and the AFNLP*. ACL; 2009, 271–279.
80. Velardi P, Faralli S, Navigli, R. OntoLearn reloaded: a graph-based algorithm for taxonomy induction. *Comput Ling* 2013, 39, 1–43.
81. Kozareva Z, Hovy E. A semi-supervised method to learn and construct taxonomies using the web. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics; 2010, 1110–1118.
82. Vossen P. Extending, trimming and fusing WordNet for technical documents. In: *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*. ACL; 2001.
83. Medelyan A, Manion S, Broekstra J, Divoli A, Huang AL, Witten IH. Constructing a focused taxonomy from a document collection. In: *Proceedings of Extended Semantic Web Conferece, ESWC*; 2013.
84. Ponzetto SP, Navigli R. Large-scale taxonomy mapping for restructuring and integrating Wikipedia. In: *Proceedings of the 21st International Joint Conference on Artificial Intelligence*. Pasadena, CA; 2009, 2083–2088.
85. Hovy E, Navigli R, Ponzetto SP. Collaboratively built semi-structured content and artificial intelligence: the story so far. *Artif Intel* 2012, 194:2–27.
86. Schreiber G, Akkermans H, Anjewierden A, de Hoog R, Shadbolt N, Van de Velde W, Wielinga B. *Knowledge Engineering and Management: The CommonKADS Methodology*. Cambridge: MIT press; 1999.
87. Lee CS, Kao YF, Kuo YH, Wang MH. Automated ontology construction for unstructured text documents. *Data Knowled Eng* 2007, 60:547–566.
88. Poon H, Domingos P. Unsupervised ontology induction from text. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics; 2010, 296–305.
89. Carlson A, Betteridge J, Kisiel B, Settles B, Hruschka Jr, ER, Mitchell TM. Toward an architecture for never-ending language learning. In: *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence*. AAAI; 2010, 4.
90. Pasca M, Lin D, Bigham J, Lifchits A, Jain A. Organizing and searching the world wide web of facts-step one: the one-million fact extraction challenge. In: *Proceedings of the National Conference on Artificial Intelligence*. MIT Press; 2006, 1400.
91. Yates A, Cafarella M, Banko M, Etzioni O, Broadhead M, Soderland S. TextRunner: open information extraction on the web. In: *Proceedings of Human Language Technologies: The Annual Conference of the NACCL: Demonstrations*. Association for Computational Linguistics; 2007, 25–26.
92. Fader A, Soderland S, Etzioni O. Identifying relations for open information extraction. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics; 2011, 1535–1545.
93. Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z. DBPedia: a nucleus for a web of open data. In: *Proceedings of the 6th International Semantic Web Conference (ISWC 2007)*. Busan, Korea; 2007, 722–735.
94. Nastase V, Michael S. Transforming Wikipedia into a large scale multilingual concept network. *Artif Intel* 2013, 194:62–85.

95. Wu F, Weld DS. Automatically refining the wikipedia infobox ontology. In: *Proceedings of the International Conference on World Wide Web*. ACM; 2008, 635–644.
96. Suchanek FM, Kasneci G, Weikum G. YAGO: a core of semantic knowledge. In: *Proceedings of the International Conference on World Wide Web*. ACM; 2007, 697–706.
97. Navigli R, Ponzetto SP. BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif Intel* 2012, 193:217–250.
98. Navigli R, Velardi P. From glossaries to ontologies: extracting semantic structure from textual definitions. In: *Proceeding of the Conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*; 2008, 71–104.
99. Gurevych I, Eckle-Kohler J, Hartmann S, Matuschek M, Meyer CM, Wirth C. UBY—a large-scale unified lexical-semantic resource based on LMF. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France; April 23–27, 2012, 580–590.
100. Hoffart J, Suchanek FM, Berberich K, Weikum G. YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia. *Artif Intel* 2012, 194:28–61.
101. Ruiz-Casado M, Alfonseca E, Castells P. Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets. *Adv Web Intel* 2005, 3528:947–950.
102. de Melo G, Weikum G. Untangling the cross-lingual link structure of Wikipedia. In: *Proceedings of the ACL*; 2010.
103. de Melo G, Weikum G. MENTA: inducing multilingual taxonomies from Wikipedia. In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. ACM; 2010, 1099–1108.
104. Maedche A, Staab S. Ontology learning for the semantic web. *Intel Syst, IEEE* 2001, 16:72–79.
105. Cimiano P, Völker J. Text2Onto. *Nat Lang Proces Inform Syst* 2005, 3513:257–271.
106. Koenderink N, van Assem M, Hulzebos J, Broekstra J, Top J. ROC: a method for proto-ontology construction by domain experts. *Semantic Web* 2008, 5367:152–166.
107. Völker J, Vrandečić D, Sure Y, Hotho A. Learning disjointness. *Semantic Web: Res Appl* 2007, 4519:175–189.
108. Merrill GH. Ontological realism: methodology or misdirection? *Appl Ontol* 2010, 5:79–108.
109. Smith B, Ceusters W. Ontological realism: a methodology for coordinated evolution of scientific ontologies. *Appl Ontol* 2010, 5:139–188.
110. Yu J, Thom JA, Tam A. Requirements-oriented methodology for evaluating ontologies. *Inform Syst* 2009, 34:766–791.
111. Yao L, Divoli A, Mayzus I, Evans JA, Rzhetsky A. Benchmarking ontologies: bigger or better? *PLoS Comput Biol* 2001, 7:e1001055.
112. Maynard D, Peters W, Li Y. Metrics for evaluation of ontology-based information extraction. In: *WWW 2006 Workshop on Evaluation of Ontologies for the Web (EON)*; 2006.
113. Smith B. From concepts to clinical reality: an essay on the benchmarking of biomedical terminologies. *J Biomed Inform* 2006, 39:288–298.

