Ontology Learning from Text: A Look Back and into the Future

WILSON WONG, RMIT University
WEI LIU and MOHAMMED BENNAMOUN, University of Western Australia

Ontologies are often viewed as the answer to the need for interoperable semantics in modern information systems. The explosion of textual information on the Read/Write Web coupled with the increasing demand for ontologies to power the Semantic Web have made (semi-)automatic ontology learning from text a very promising research area. This together with the advanced state in related areas, such as natural language processing, have fueled research into ontology learning over the past decade. This survey looks at how far we have come since the turn of the millennium and discusses the remaining challenges that will define the research directions in this area in the near future.

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Abstracting methods*; I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods—*Semantic networks*; *representation languages*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Language parsing and understanding*; *text analysis*

General Terms: Algorithms, Languages

Additional Key Words and Phrases: Ontology learning, term recognition, concept discovery, semantic relation acquisition, application of ontologies

ACM Reference Format:
Wilson, W., Liu, W., and Bennamoun, M. 2012. Ontology learning from text: A look back and into the future. ACM Comput. Surv. 44, 4, Article 20 (August 2012), 36 pages.
DOI = 10.1145/2333112.2333115 http://doi.acm.org/10.1145/2333112.2333115

1. INTRODUCTION

Advances in areas such as natural language processing, information retrieval, machine learning, data mining, and knowledge representation have been fundamental in our quest for means to make sense of an ever growing body of textual information in electronic forms, known simply as *information* from here on. The intermingling of techniques from these areas has enabled us to extract and represent facts and patterns for improving the management, access, and interpretability of information. However, it was not until the turn of the millennium with the *Semantic Web* dream [Maedche and Staab 2001] and the explosion of information due to the *Read/Write Web* that the need for a systematic body of study in large-scale extraction and representation of facts and patterns became more obvious. Over the years, that realization gave rise to a research area now known as *ontology learning from text* which aims to turn facts and patterns from an ever growing body of information into shareable high-level constructs

This work was partly funded by the Australian Endeavour International Postgraduate Research Scholarship. This work began when W. Wong was with the University of Western Australia.

Authors' addresses: W. Wong, RMIT University, Melbourne Victoria 3000, Australia; email: wilson.wong@rmit.edu.au; W. Liu, University of Western Australia, Crawley WA 6009, Australia; email: wei@csse.uwa.edu.au; M. Bennamoun, University of Western Australia, Crawley WA 6009, Australia; email: bennamou@csse.uwa.edu.au.

DOI 10.1145/2333112.2333115 http://doi.acm.org/10.1145/2333112.2333115

for enhancing everyday applications (e.g., Web search) and enabling intelligent systems (e.g., Semantic Web).

Ontologies are effectively formal and explicit specifications in the form of concepts and relations of shared conceptualizations [Gruber 1993]. Ontologies may contain axioms for validation and enforcing constraints. There has always been a subtle confusion or controversy regarding the difference between an ontology and a knowledge base. In an attempt to draw a line between these two structures, consider the loosely applicable analogy where ontologies are cupcake molds and knowledge bases are the actual cupcakes of assorted colours, tastes, and so on. Ontologies, in this sense, represent the intensional aspect of a domain for governing the way the corresponding knowledge bases (i.e., extensional aspect) are populated [Buitelaar et al. 2005]. In other words, every knowledge base has to be committed to a conceptualization, whether implicitly or explicitly. This conceptualization is what we refer to as ontologies [Gruber 1993]. With this in mind, knowledge bases can be created by extracting the relevant instances from information to populate the corresponding ontologies, a process known as *ontology population* or *knowledge markup*. Ontology learning from text is then essentially the process of deriving high-level concepts and relations as well as the occasional axioms from information to form an ontology.

Ontology learning has benefited from the adoption of established techniques from the related areas just discussed. Aside from the inherent challenges of processing natural language, one of the remaining obstacles preventing the large-scale deployment of ontology learning systems is the bottleneck in handcrafting structured knowledge sources (e.g., dictionaries, taxonomies, knowledge bases) [Cullen and Bryman 1988] and training data (e.g., annotated text corpora). It is gradually becoming apparent that in order to minimize human efforts in the learning process and to improve the scalability and robustness of the system, static and expert crafted resources may no longer be adequate. Recognizing this, an increasing amount of research effort is gradually being directed towards harnessing the collective intelligence on the Web in the hopes of addressing this one major bottleneck. At the same time, as with many fields before ontology learning, the process of maturing has triggered a mounting awareness of the actual intricacies involved in automatically discovering concepts, relations, and even axioms. This gives rise to the question of whether the ultimate goal of achieving full-fledged formal ontologies automatically can be achieved. While certain individuals dwell on the question, many others move on with a more pragmatic goal, which is to focus on learning lightweight ontologies first and extend them later if possible. With high hopes and achievable aims, we are seeing a gradual rise in the adoption of ontologies across many domains that require knowledge engineering, in particular, interoperability of semantics in their applications (e.g., document retrieval [Castells et al. 2007], image retrieval [Hyvonen et al. 2003], bioinformatics [Baker et al. 2007], manufacturing [Cho et al. 2006], industrial safety [Abou-Assali et al. 2007], law [Volker et al. 2008], environment [Raskin and Pan 2005], disaster management [Klien et al. 2006], e-Government [Kayed et al. 2010], e-Commerce [Liu et al. 2008], and tourism [Park et al. 2009]).

This article provides a comprehensive review of (1) the process of ontology learning in general, (2) the smorgasbord of techniques for ontology learning, (3) the lack of common evaluation platforms, (4) seven prominent ontology learning systems, (5) the progress to date, and (6) the outstanding challenges. Section 2 looks at five surveys conducted in the past decade. Section 3 contains an introduction to ontologies and the process of ontology learning. The definition of an ontology is first provided, followed by a discussion on the differences between lightweight versus formal ontologies. The process of ontology learning is then described, with a focus on the types of output. The section moves on to describe the most commonly used techniques in ontology learning which are borrowed from established areas, such as natural language processing, information retrieval, data mining and knowledge representation. Methodologies and benchmarks used for evaluating ontology learning techniques are also discussed. Section 4 then goes through in detail seven of the more established ontology learning systems in the past decade. Recent advances including new techniques and emerging data sources for ontology learning are then presented in Section 5. We bring this articleto an end in Section 6 with a summary of all sections and a discussion of the remaining challenges in ontology learning.

2. A LOOK AT PREVIOUS SURVEYS

Before a discussion on ontology learning in general and seven prominent systems, we take a brief look at five previous independent surveys since 2000. The first is a 2003 report by the OntoWeb Consortium [Gomez-Perez and Manzano-Macho 2003], a body funded by the Information Society Technologies Programme of the Commission of the European Communities. This survey listed 36 approaches for ontology learning from text. The important findings presented by this review paper are (1) the lack of a detailed methodology that guides the ontology learning process from text; (2) no fully automated system for ontology learning and many require the involvement of users in order to extract concepts and relations from annotated corpora; and (3) a need for a general approach for evaluating the accuracy of ontology learning and for comparing the results produced by different systems.

The second survey, released about the same time as the OntoWeb Consortium survey, was performed by Shamsfard & Barforoush [2003]. The authors claimed to have studied over 50 different approaches before selecting and including seven prominent ones in their survey. The main focus of the review was to introduce a framework for comparing ontology learning approaches. The approaches included in the review merely served as test cases to be fitted into the framework. Consequently, the review provided an extensive coverage of the state of the art of the relevant techniques but was limited in terms of discussions on the underlying problems and future outlook. The review arrived at the following list of problems: (1) much work has been conducted on discovering taxonomic relations, while non-taxonomic relations were given less attention; (2) research into axiom learning is unexplored; (3) the focus of most research is on building domain ontologies and most systems were designed to make heavy use of domain-specific patterns and static background knowledge with little regard to the portability of the systems across different domains; (4) current ontology learning systems are evaluated within the confinement of their domains, and finding a formal standard method of evaluating ontology learning systems remains an open problem; and (5) most systems are either semi-automated or tools for supporting domain experts in curating ontologies.

Third, Ding and Foo [2002] presented a survey of 12 major ontology learning projects. The authors wrapped up their survey with the following findings: (1) input data are mostly structured, and learning from free texts remains within the realm of research; (2) the task of discovering relations is very complex and a difficult problem to solve, and it has turned out to be the main impediment to the progress of ontology learning; and (3) the techniques for discovering concepts have reached a certain level of maturity.

Fourth, Buitelaar et al. [2005] brought together the research published at two workshops on ontology learning and knowledge acquisition in 2004. The editors organized the ten papers included in their book into methodologies, evaluation methods, and application scenarios. The editors also popularized the use of the phrase *“ontology learning layer cake”* to describe the different subtasks involved in ontology learning. The editors presented three main observations: (1) there is a need for more research in the axiom extraction subtask; (2) the importance of a common evaluation platform for promoting progress in the area of ontology learning; and (3) the gradual departure from small static text collections to Web resources for ontology learning.

Last, Zhou [2007] published a brief survey looking at several outstanding challenges in the area. In this paper, the author also proposed a learning-oriented model for the development of ontology. More importantly, the five issues highlighted by the author which are relevant to our discussion include: (1) the importance of representation in the development of ontologies; (2) the involvement of humans in ontology learning remains highly necessary and desirable; (3) the need for common benchmarks for evaluating ontologies from a variety of perspectives (e.g., domains); (4) while there is some progress on acquiring generic relations, more work is required in the discovery of fine-grained associations; and (5) more research effort is required to make existing techniques operational on cross-domain text on a Web-scale, and this includes the challenge of acquiring the knowledge necessary for learning ontologies (i.e., knowledge acquisition bottleneck). The author agreed that expert-curated domain knowledge is no longer adequate and highlighted the fact that researchers are turning to other sources on the Web for the content of ontologies.

A closer look into the five survey papers revealed a consensus on several aspects of ontology learning that required more work. The main realizations that remain valid over the past decade are that (1) the fully automatic learning of ontologies may not be possible, (2) a lack of common evaluation platforms for ontologies is evident, and (3) the discovery of relations between concepts, especially fine-grained ones, requires more work. Not surprisingly, one additional conclusion that can be drawn from the more recent literature (i.e., second part of the decade) is (4) the increase in interest in harnessing the Web to address the knowledge acquisition bottleneck and to make ontology learning operational on a Web-scale. The validity of these four conclusions from the five surveys will become evident as we look into several prominent systems and recent advances in ontology learning in Sections 4 and 5.

3. ONTOLOGY LEARNING FROM TEXT

Ontology learning from text is the process of identifying terms, concepts, relations, and optionally, axioms from textual information and using them to construct and maintain an ontology. Techniques from established fields, such as information retrieval, data mining, and natural language processing, have been fundamental in the development of ontology learning systems.

3.1. Lightweight versus Formal Ontologies

Ontologies can be thought of as directed graphs consisting of *concepts* as nodes and *relations* as the edges between the nodes. A concept is essentially a mental symbol often realized by a corresponding lexical representation (i.e., natural language name). For instance, the concept *"food"* denotes the set of all substances that can be consumed for nutrition or pleasure. In Information Science, an ontology is a *"formal, explicit specification of a shared conceptualisation"* [Gruber 1993]. This definition imposes the requirement that the names of concepts and how the concepts are related to one another have to be explicitly expressed and represented using formal languages, such as Web Ontology Language (OWL).[1] An important benefit of a formal representation is the ability to specify *axioms* for reasoning in order to determine validity and to define constraints in ontologies. Moreover, a formal ontology is natural language independent or, in other words, does not contain lexical knowledge [Hjelm and Volk 2011].

As research into ontology learning progresses, the definition of what constitutes an ontology evolves. The extent of relational and axiomatic richness and the formality of representation eventually gave rise to a spectrum of ontology kinds [Uschold and Gruninger 2004]. Figure 1, which was adapted from Giunchiglia and Zaihrayeu [2007],

[1] `http://www.w3.org/TR/owl-ref/`.

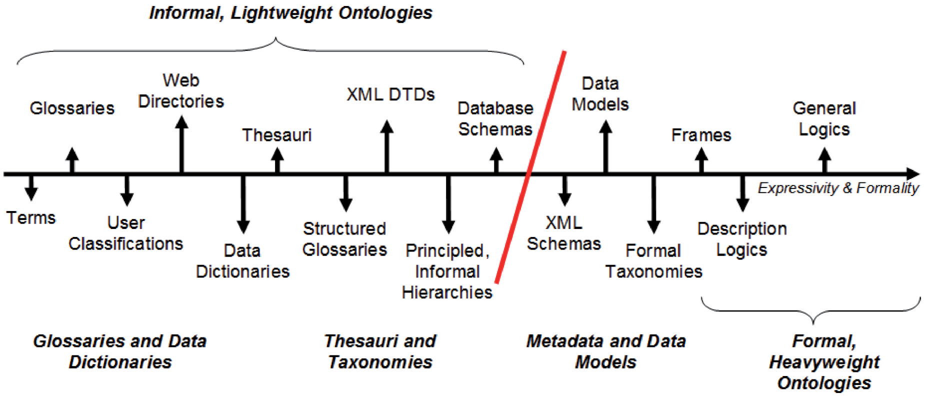


Fig. 1. The spectrum of ontology kinds.

illustrates this spectrum. At one end of the spectrum, we have ontologies that make little or no use of axioms, referred to as *lightweight ontologies* [Giunchiglia and Zaihrayeu 2007]. At the other end, we have *heavyweight ontologies* [Furst and Trichet 2006] that make intensive use of axioms for specification. Glossaries and dictionaries can be referred to collectively as controlled vocabularies. A controlled vocabulary is a list of terms that have been enumerated explicitly and maintained or regulated by independent authorities. Theoretically, terms in a controlled vocabulary should be defined in a way to minimize or avoid ambiguity and redundancy. A taxonomy, on the other hand, is a controlled vocabulary organized into a hierarchical or parent-child structure. A thesaurus is similar to a taxonomy, with the addition of more relationships beyond hierarchical.

Ontologies are fundamental to the success of the Semantic Web, as they enable software agents to exchange, share, reuse, and reason about concepts and relations using axioms. In the words of Berners-Lee et al. [2001], “For the semantic web to function, computers must have access to structured collections of information and sets of inference rules that they can use to conduct automated reasoning.” However, the truth remains that the automatic learning of axioms is not an easy task. Despite certain success, many ontology learning systems are still struggling with the basics of extracting terms and relations [Furst and Trichet 2006]. For this reason, the majority of ontology learning systems out there that claim to learn ontologies are, in fact, creating lightweight ontologies. At the moment, lightweight ontologies appear to be the most common type of ontologies in a variety of Semantic Web applications (e.g., knowledge management, document retrieval, communities of practice, data integration) [Davies et al. 2003; Fluit et al. 2003].

3.2. Outputs and Tasks in Ontology Learning

There are five types of output in ontology learning, namely, *terms*, *concepts*, *taxonomic relations*, *non-taxonomic relations*, and *axioms*. Some researchers [Buitelaar et al. 2005] refer to this as the *ontology learning layer cake*. To obtain each output, certain tasks have to be accomplished, and the techniques employed for each task may vary between systems. This view of output-task relation that is independent of any implementation details promotes modularity in designing and implementing ontology learning systems. Figure 2, initially introduced in Wong [2009], shows the outputs, the corresponding tasks, and the plethora of typically employed techniques. Each output is a prerequisite for obtaining the next output, as shown in the figure. Terms are used to form concepts which in turn are organized according to relations. Relations can be further generalized to produce axioms. The solid arrows are used to relate techniques

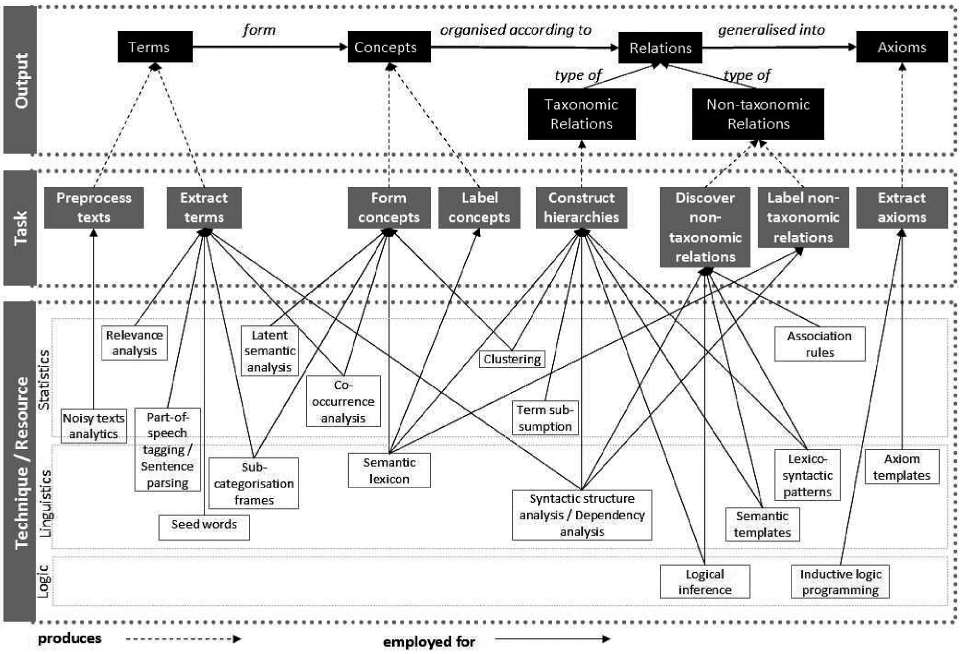


Fig. 2. An overview of the outputs, tasks, and common techniques for ontology learning. The tasks are connected to the outputs using dotted arrows which represent the association *“produces,”* while the solid arrows refer to the *“employed for”* association to relate techniques and resources to tasks.

to tasks (i.e., `technique X employed for task Y`), while the dotted arrows indicate the connections between tasks and outputs (i.e., `task X produces output Y`).

Terms are the most basic building blocks in ontology learning. Terms can be simple (i.e., single word) or complex (i.e., multi word), and are considered as lexical realizations of everything important and relevant to a domain. The main tasks associated with terms are to *preprocess texts* and *extract terms*. The preprocessing task ensures that the input texts are in a format supported by the ontology learning system. Some of the techniques relevant to preprocessing include noisy text analytics and the extraction of relevant contents from webpages (i.e., boilerplate removal). The extraction of terms, known as *term extraction* or *keyphrase extraction* [Medelyan and Witten 2005], typically begins with tokenization or part-of-speech tagging to break texts into smaller constituents. Statistical or probabilistic measures are then used to determine the collocational stability of a noun sequence to form a term, also known as *unithood*, and the relevance or specificity of a term with respect to a domain, also known as *termhood*.

Concepts can be abstract or concrete, real or fictitious. Broadly speaking, a concept can be anything about which something is said. Concepts are formed by grouping similar terms. The main tasks are therefore to *form concepts* and *label concepts*. The task of forming concepts involves discovering the variants of a term and grouping them together. Term variants can be determined using predefined background knowledge, syntactic structure analysis, or through clustering based on some similarity measures. Syntactic structure analysis, for instance, uses the common head such as *“tart”* of complex terms to form a unifying concept to encompass the corresponding longer strings *“egg tart”*, *“French apple tart”*, and *“chocolate tart”*. If labels are required for the concepts, existing background knowledge, such as WordNet, may be used to find the name of the nearest common ancestor.

Relations are used to model the interactions between the concepts in an ontology. There are two types of relations, namely, *taxonomic relations* and *non-taxonomic relations*. The main task that involves taxonomic relations is to *construct hierarchies*. Organizing concepts into a hierarchy requires the discovery of `is-a` relations (i.e., hypernym/hyponym) [Cimiano et al. 2004], and hence, some researchers may also refer to this task as *extracting taxonomic relations*. Hierarchy construction can be performed in various ways, such as using predefined relations from existing background knowledge, using statistical subsumption models, relying on semantic similarity between concepts, and utilizing linguistic and logical rules or patterns. Non-taxonomic relations are the interactions between concepts (e.g., meronymy, thematic roles, attributes, possession, and causality) other than hypernymy. The less explicit and more complex use of words for specifying relations other than hypernymy causes the tasks of *discovering non-taxonomic relations* and *labeling non-taxonomic relations* to be more challenging. Discovering and labeling non-taxonomic relations are mainly reliant on the analysis of syntactic structures and dependencies. In this aspect, verbs are taken as good indicators for non-taxonomic relations, and help from domain experts may be required to label such relations.

Lastly, axioms are propositions or sentences that are always taken as true. Axioms act as a starting point for deducing other truth, verifying correctness of existing ontological elements, and defining constraints. The task involved here is of *discovering axioms*. The task of learning axioms involves the generalization or deduction of a large number of known relations that satisfy certain criteria.

3.3. Techniques for Ontology Learning

Many proven techniques from established fields, such as information retrieval, machine learning, data mining, natural language processing, as well as knowledge representation and reasoning, have all contributed to the progress in ontology learning over the past decade. Information retrieval provides various algorithms for analyzing associations between concepts in texts using vectors, matrices [Fortuna et al. 2005], and probabilistic theorems [Yang and Calmet 2005]. On the other hand, machine learning and data mining provides ontology learning the ability to extract rules and patterns out of massive datasets in a supervised or unsupervised manner based on extensive statistical analysis. Natural language processing provides the tools for analyzing natural language text on various language levels (e.g., morphology, syntax, semantics) to uncover concept representations and relations through linguistic cues. Knowledge representation and reasoning enables the ontological elements to be formally specified and represented such that new knowledge can be deduced. The techniques employed by different systems may vary depending on the tasks to be accomplished. The techniques can generally be classified into *statistics-based*, *linguistics-based*, *logic-based*, or *hybrid*. The purpose of this classification is to compartmentalize discussions. It is not our intention to contrast the different techniques to decide on which is better. In reality, the hybrid approach is mainly used in existing studies. Figure 2 illustrates the various commonly used techniques and their associations to the different tasks. *Bootstrapping* is a popular approach used to kickstart the construction of ontologies based on some user-provided resources, also known as *seeds*. A combination of these preceding techniques is then used to extend the seeds. Brewster et al. [2002] described a methodology for constructing an ontology using a text corpus and an existing or a sketch of a preliminary ontology. Liu et al. [2005] presented a semi-automatic approach to extending and refining seed ontologies by mining webpages on online media sites.

3.3.1. Statistics-Based Techniques. The various statistics-based techniques for accomplishing the tasks in ontology learning are mostly derived from information retrieval, machine learning, and data mining. The lack of consideration for the underlying semantics and relations between the components of a text makes statistics-based techniques more prevalent in the early stages of ontology learning, such as term extraction and hierarchy construction. Some of the common techniques include *clustering* [Wong et al. 2007], *latent semantic analysis* [Turney 2001], *cooccurrence analysis* [Budanitsky 1999], *term subsumption* [Fotzo and Gallinari 2004], *contrastive analysis* [Velardi et al. 2005], and *association rule mining* [Srikant and Agrawal 1997]. The main idea behind these techniques is that the (co-)occurrence of lexical units[2] in samples often provides a reliable estimate about their semantic identity to enable the creation of higher-level entities.

—In clustering, some measure of similarity is employed to assign terms into groups for discovering concepts or constructing hierarchy [Linden and Piitulainen 2004]. The process of clustering can either begin with individual terms or concepts, grouping the most related ones (i.e., agglomerative clustering), or begin with all terms or concepts and dividing them into smaller groups to maximize within group similarity (i.e., divisive clustering). Some of the major issues in clustering are working with high-dimensional data and feature extraction and preparation for similarity measurement. This gave rise to a class of featureless similarity measures based solely on the cooccurrence of words in large text corpora. The Normalised Web Distance (NGD)[3] is one example [Vitanyi et al. 2009]. For clarification, there are in fact two types of similarity, namely, *paradigmatic similarity* and *syntagmatic similarity*. Two terms are paradigmatically similar if they are substitutable for one another in a particular context (e.g., "apple" and "orange"). Syntagmatic similarity, on the other hand, refers to the association between terms related through significant cooccurrence (e.g., "cut" and "knife"). From these examples, we can say that "apple" and "orange" are similar, as in they are fruits, while "cut" and "knife" are related (definitely not similar), since they are used for preparing food. Technically, we would refer to the former as *semantic similarity*, while the latter as *semantic relatedness*. However, throughout this article, no such distinction will be made.

—Relying on raw data to measure similarity may lead to data sparseness [Buitelaar et al. 2005]. Originally applied to indexing documents in information retrieval, latent semantic analysis and other related approaches based on dimension-reduction techniques are applied on term-document matrices to overcome the problem [Landauer et al. 1998]. More importantly, the inherent relations between terms can be revealed by applying correlation measures on the dimensionally reduced matrix, leading to the formation of concepts.

—Cooccurrence analysis attempts to identify lexical units that tend to occur together for purposes ranging from extracting related terms to discovering implicit relations between concepts. Cooccurrence can appear in many forms, such as on the phrasal level (e.g., "black jack", "governed by") or through common associations (e.g., "Steve" and "Apple"). The cooccurrence of a sequence of words (i.e., multi-word expression) beyond chance within a well-defined unit (e.g., phrase, sentence) is called a collocation. Cooccurrence measures are used to determine the association strength between terms or the constituents of terms. Some of the popular measures include dependency measures (e.g., mutual information [Church and Hanks 1990]), log-likelihood ratios [Resnik 1999] (e.g., chi-square test), rank correlations (e.g., Pearson's and Spearman's

[2] A single word or chain of words that are the basic elements of a vocabulary.

[3] The original distance measure based on the Google search engine is known as the normalized Google distance.

coefficient [Strehl 2002]), and similarity measures (e.g., cosine measures [Senellart and Blondel 2003], Kullback-Leiber divergence [Maedche et al. 2002]).

—In term subsumption, the conditional probabilities of the occurrence of terms in documents are employed to discover hierarchical relations between them [Fotzo and Gallinari 2004]. A term subsumption measure is used to quantify the extent of a term x being more general than another term y. The higher the subsumption value, the more general term x is with respect to y.

—The extent of occurrence of terms in individual documents and in text corpora is employed for relevance analysis. Some of the common relevance measures from information retrieval include the term frequency-inverse document frequency (TF-IDF) [Salton and Buckley 1988] and its variants, and others based on language modeling [Ponte and Croft 1998] and probability [Fuhr 1992]. Contrastive analysis [Basili et al. 2001] is a kind of relevance analysis based on the heuristic that general language-dependent phenomena should spread equally across different text corpora, while special-language phenomena should portray odd behaviours.

—Given a set of concept pairs, association rule mining is employed to describe the associations between the concepts at the appropriate level of abstraction [Jiang et al. 2007]. In the example by Maedche and Staab [2000a], given the already known concept pairs {*chips, beer*} and {*peanuts, soda*}, association rule mining is then employed to generalize the pairs to provide {*snacks, drinks*}. The key to determining the degree of abstraction in association rules is provided by user-defined thresholds, such as confidence and support.

3.3.2. Linguistics-Based Techniques and Resources. Linguistics-based techniques are applicable to almost all tasks in ontology learning and are mainly dependent on natural language processing tools. Some of the techniques include *part-of-speech tagging*, *sentence parsing*, *syntactic structure analysis*, and *dependency analysis*. Other techniques rely on the use of *semantic lexicon*, *lexico-syntactic patterns*, *semantic templates*, *subcategorization frames*, and *seed words*.

—Part-of-speech tagging and sentence parsing provide the syntactic structures and dependency information required for further linguistic analysis in order to uncover terms and relations. Some examples of part-of-speech tagger are Brill Tagger [Brill 1992] and TreeTagger [Schmid 1994]. Principar [Lin 1994], Minipar [Lin 1998], and Link Grammar Parser [Sleator and Temperley 1993] are amongst the few common sentence parsers. Other more comprehensive toolkits for natural language processing include General Architecture for Text Engineering (GATE) [Cunningham et al. 2002], and Natural Language Toolkit (NLTK) [Bird et al. 2008]. Despite the placement under the linguistics-based category, certain parsers are built on statistical parsing systems. For instance, the Stanford Parser [Klein and Manning 2003] is a lexicalized probabilistic parser.

—Syntactic structure analysis and dependency analysis examines syntactic and dependency information to uncover terms and relations at the sentence level [Sombatsrisomboon et al. 2003]. In syntactic structure analysis, words and modifiers in syntactic structures (e.g., noun phrases, verb phrases, and prepositional phrases) are analyzed to discover potential terms and relations. For example, `ADJ-NN` or `DT-NN` can be extracted as potential terms while ignoring phrases containing other part of speech, such as verbs. In particular, the head-modifier principle has been employed extensively to identify complex terms related through hyponymy, with the heads of the terms assuming the hypernym role [Hippisley et al. 2005]. In dependency analysis, grammatical relations, such as subject, object, adjunct, and complement, are used for determining more complex relations [Gamallo et al. 2002; Ciaramita et al. 2005].

- Semantic lexicons are a popular resource in ontology learning. They can either be general, such as WordNet [Miller et al. 1990], or domain specific, such as the Unified Medical Language System (UMLS) [Lindberg et al. 1993]. Many of the works related to the use of WordNet can be found in the areas of lexical acquisitions [O'Hara et al. 1998], word sense disambiguation [Vronis and Ide 1998; Lesk 1986], as well as similarity measurement [Pedersen et al. 2004]. Semantic lexicons offer easy access to a large collection of predefined concepts and relations. Concepts from semantic lexicon are organized in sets of similar words (i.e., synsets). These synonyms are employed for discovering variants of terms [Turcato et al. 2000] for forming concepts. The associations defined in lexicons such as hypernym-hyponym (i.e., parent-child relation) and meronym-holonym (i.e., part-whole relation), on the other hand, have been proven useful to the tasks of taxonomic and non-taxonomic relation extraction.
- The use of lexico-syntactic patterns was proposed by Hearst [1998] and has been employed to extract hypernyms [Sombatsrisomboon et al. 2003] and meronyms. Lexico-syntactic patterns capture hypernymy relations using patterns such as `NP such as NP, NP,..., and NP`. For extracting meronyms, patterns such as `NP is part of NP` can be useful. The use of patterns provide reasonable precision, but the recall is low [Buitelaar et al. 2005]. Due to the cost and time involved in manually producing such patterns, efforts [Snow et al. 2005] have been taken to study the possibility of learning them. Semantic templates [Spiliopoulou et al. 2004; Vargas-Vera et al. 2001] are similar to lexico-syntactic patterns in terms of their purpose. However, semantic templates offer more detailed rules and conditions for extractinf not only taxonomic relations but also complex non-taxonomic relations.
- In linguistic theory, the subcategorization frame [Agustini et al. 2001; Gamallo et al. 2003] of a word is the number and kinds of other words that it selects when appearing in a sentence. For example, in the sentence "Joe wrote a letter", the verb "write" selects "Joe" and "letter" as its subject and object, respectively. In other words, "Person" and "written-communication" are the *restrictions of selection* for the subject and object of the verb "write". The restrictions of selection extracted from parsed texts can be used in conjunction with clustering techniques to discover concepts [Faure and Nedellec 1998a].
- The use of seed words (i.e., seed terms) [Yangarber et al. 2000] is a common practice in many systems for guiding a wide range of tasks in ontology learning. Seed words provide good starting points for the discovery of additional terms relevant to that particular domain [Hwang 1999]. Seed words are also used to guide the automatic construction of text corpora from the Web [Baroni and Bernardini 2004].

3.3.3. Logic-Based Techniques and Resources. Logic-based techniques are the least common in ontology learning and are mainly adopted for more complex tasks involving relations and axioms. Logic-based techniques have connections with advances in knowledge representation and reasoning and in machine learning. The two main techniques employed are *inductive logic programming* [Lavrac and Dzeroski 1994; Zelle and Mooney 1993] and *logical inference* [Shamsfard and Barforoush 2004].

- In inductive logic programming, rules are derived from existing collection of concepts and relations which are divided into positive and negative examples. The rules proves all the positive and none of the negative examples. In an example by Oliveira et al. [2001], induction begins with the first positive example "tigers have fur". With the second positive example "cats have fur", a generalisation of "felines have fur" is obtained. Given the third positive example "dogs have fur", the technique will attempt to generalize that "mammals have fur". When encountered with a negative example "humans do not have fur", then the previous generalization will be dropped, giving only "canines and felines have fur."

—In logical inference, implicit relations are derived from existing ones using rules such as transitivity and inheritance. Using the classic example, given the premises “Socrates is a man” and “All men are mortal”, we can discover a new attribute relation stating that “Socrates is mortal”. Despite the power of inference, the possibilities of introducing invalid or conflicting relations may occur if the design of the rules is not complete. Consider the example in which “human eats chicken” and “chicken eats worm” yields a new relation that is not valid. This happened because the intransitivity of the relation “eat” was not explicitly specified in advance.

3.4. Evaluation of Ontology Learning Techniques

Evaluation is an important aspect of ontology learning, just like any other research areas. Evaluation allows individuals who use ontology learning systems to assess the resulting ontologies and to possibly guide and refine the learning process. An interesting aspect about evaluation in ontology learning, as opposed to information retrieval and other areas, is that ontologies are not an end product but, rather, a means to achieving some other tasks. In this sense, an evaluation approach is also useful to assist users in choosing the best ontology that fits their requirements when faced with a multitude of options.

In document retrieval, the object of evaluation is documents and how well systems provide documents that satisfy user queries, either qualitatively or quantitatively. However, in ontology learning, we cannot simply measure how well a system constructs an ontology without raising more questions. For instance, is the ontology good enough? If so, with respect to what application? An ontology is made up of different layers, such as terms, concepts, and relations. If an ontology is inadequate for an application, then which part of the ontology is causing the problem? Considering the intricacies of evaluating ontologies, a myriad of evaluation approaches have been proposed in the past few years. Generally, these approaches can be grouped into one of the following three main categories depending on the kind of ontologies that are being evaluated and the purpose of the evaluation [Brank et al. 2005].

—The first approach evaluates the adequacy of ontologies in the context of other applications. For example, in the case of an ontology designed to improve the performance of document retrieval, we may collect some sample queries and determine if the documents retrieved are actually more relevant when the ontology is used. Porzel and Malaka [2004] evaluated the use of ontological relations in the context of speech recognition. The output from the speech recognition system is compared with a gold standard generated by humans. In particular, Lozano-Tello et al. [2003] proposed a methodology which allows users to assess how well an ontology meets their systems’ requirements. The choice of an objective measure for such an evaluation depends on the task. In our example of a document retrieval system, conventional measures in information retrieval, such as F-score, may be used. This approach of evaluation is also known as *task-based evaluation* [Dellschaft and Staab 2008].

—The second approach uses domain-specific data sources to determine to what extent the ontologies are able to cover the corresponding domain. For instance, Brewster et al. [2004] described a number of methods for evaluating the ‘fit’ between an ontology and the domain knowledge in the form of text corpora. In this approach, natural language processing (e.g., latent semantic analysis, clustering) or information extraction (e.g., named-entity recognition) techniques are used to analyze the content of the corpus and identify terms. The terms are then compared against the content of the ontology to be evaluated. This approach of evaluation is also known as *corpus-based evaluation* [Dellschaft and Staab 2008].

—The third approach, also known as *criteria-based evaluation* [Dellschaft and Staab 2008], assesses ontologies by determining how well they adhere to a set of criteria. For example, one may set as part of the criteria the average number of terms that were aggregated to form a concept in an ontology. This criterion may be used to realize the perception that the more variants of a term used to form a concept, the more fully encompassing or complete the concept is.

Due to the complex nature of ontologies, evaluation approaches can also be distinguished by the layers of an ontology (e.g., term, concept, relation) they evaluate [Porzel and Malaka 2004]. More specifically, evaluations can be performed to assess the (1) correctness at the terminology layer, (2) coverage at the conceptual layer, (3) wellness at the taxonomy layer, and (4) adequacy of the non-taxonomic relations.

The focus of evaluation at the terminology layer is to determine if the terms used to identify domain-relevant concepts are included and correct. Some form of lexical reference or benchmark is typically required for evaluation in this layer. Typical *precision* and *recall* measures from information retrieval are used together with exact matching or edit distance [Maedche and Staab 2002] to determine performance at the terminology layer. The lexical precision and recall reflect how good the extracted terms cover the target domain. *Lexical recall (LR)* measures the number of relevant terms extracted ($e_{relevant}$) divided by the total number of relevant terms in the benchmark ($b_{relevant}$), while *lexical precision (LP)* measures the number of relevant terms extracted ($e_{relevant}$) divided by the total number of terms extracted (e_{all}). LR and LP are defined as the following [Sabou et al. 2005].

$$LP = \frac{e_{relevant}}{e_{all}}, \tag{1}$$

$$LR = \frac{e_{relevant}}{b_{relevant}}. \tag{2}$$

The precision and recall measure also can be combined to compute the corresponding F$_\beta$-score. The general formula for non negative real β is

$$F_\beta = \frac{(1+\beta^2)(precision \times recall)}{\beta^2 \times precision + recall}. \tag{3}$$

Evaluation measures at the conceptual level are concerned with whether the desired domain-relevant concepts are discovered or otherwise. *Lexical overlap (LO)* measures the intersection between the discovered concepts (C_d) and the recommended concepts (C_m). LO is defined as

$$LO = \frac{|C_d \cap C_m|}{|C_m|}. \tag{4}$$

Ontological improvement (OI) and *ontological loss (OL)* are two additional measures to account for newly discovered concepts that are absent from the benchmark and for concepts which exist in the benchmark but were not discovered, respectively. They are defined as the following [Sabou et al. 2005].

$$OI = \frac{|C_d \backslash C_m|}{|C_m|}, \tag{5}$$

$$OL = \frac{|C_m \backslash C_d|}{|C_m|}. \tag{6}$$

Evaluations at the taxonomy layer are more complicated. Performance measures for the taxonomy layer are typically divided into local and global [Dellschaft and Staab 2006]. The similarity of the concepts' positions in the learned taxonomy and in the benchmark is used to compute the local measure. The global measure is then derived by averaging the local scores for all concept pairs. One of the few measures for the taxonomy layer is the *taxonomic overlap (TO)* [Maedche and Staab 2002]. The computation of the global similarity between two taxonomies begins with the local overlap of their individual terms. The *semantic cotopy*, that is, the set of all super- and subconcepts of a term, varies depending on the taxonomy. The local similarity between two taxonomies given a particular term is determined based on the overlap of the term's semantic cotopy. The global taxonomic overlap is then defined as the average of the local overlaps of all the terms in the two taxonomies. The same idea can be applied to compare adequacy non-taxonomic relations.

4. AN OVERVIEW OF PROMINENT ONTOLOGY LEARNING SYSTEMS

After a look at some previous surveys and some background on ontology learning in Sections 2 and 3, we now move on to examine the techniques used by seven prominent ontology learning systems and the evaluation of these techniques. The discussion of each system in the following seven sections is structured as follows. We first provide an overview of the system in terms of its developers, the motivation behind the system, and its application domains. We then elaborate on the techniques employed by each system in terms of the corresponding tasks to be achieved, as summarized in Figure 2. We end each section with a discussion on the evaluations performed on the system. As the name of this section partly indicates, these seven systems were chosen mainly for their wide adoption or popularity, their comprehensiveness in regard to the number of ontology learning tasks and outputs supported, or the recency of the work. Text-to-Onto [Cimiano and Staab 2005], for instance, is featured in this section despite its age due to its significance to a wide range of researchers as well as practitioners for purposes ranging from e-Learning [Hatala et al. 2009] and e-Government [Kayed et al. 2010] to applications in the legal domain [Volker et al. 2008]. OntoGain, on the other hand, is included due to its recency to the community, which will act as an excellent yardstick to examine the progress of ontology learning systems over the past ten years. Table I provides a summary of the seven systems reviewed in this section. A critical analysis of these systems is included in Section 6.

4.1. ASIUM

ASIUM [Faure and Poibeau 2000; Faure and Nedellec 1999, 1998b] is a semi-automated ontology learning system that is part of an information extraction infrastructure called INTEX, by the Laboratoire d'Automatique Documentaire et Linguistique de l'Universite de Paris 7. The aim of this approach is to learn semantic knowledge from texts and use the knowledge for the expansion (i.e., portability from one domain to the other) of INTEX. ASIUM uses linguistics and statistics-based techniques to perform its ontology learning tasks as described next:

—*Preprocess texts and discover subcategorization frames.* Sentence parsing is applied on the input text using functionalities provided by a sentence parser called SYLEX [Constant 1995]. SYLEX produces all interpretations of parsed sentences, including attachments of noun phrases to verbs and clauses. Syntactic structure and dependency analysis is performed to extract instantiated subcategorization frames in the form of `<verb><syntactic_role|preposition:head_noun>`*, where the wildcard character (*) indicates the possibility of multiple occurrences.

Table I. A Summary of the Outputs Supported, Techniques Used, and Evaluations Performed for the Seven Systems Included

System	Output					Technique / Resource			Evaluation Metrics
	Terms	Concepts	Taxonomic relations	Non-taxonomic relations	Axioms	Statistics-based	Linguistics-based	Logic-based	
ASIUM (2000)	√						Sentence parsing, Syntactic structure analysis, Subcategorization frames		Precision measure for term extraction
		√							
			√			Aglomerative Clustering			
Text-to-Onto (2000)	√						Part-of-speech tagging, Sentence parsing, Syntactic structure analysis		F-measure and generic relations learning accuracy (RLA) for non-taxonomic relation extraction
		√				Co-occurrence analysis	Concepts from domain lexicon		
			√			Aglomerative Clustering	Hypernyms from WordNet, Lexico-syntactic patterns		
				√		Association rule mining			
TextStorm/ Clouds (2001)	√						Part-of-speech tagging using WordNet, Syntactic structure analysis, Anaphora resolution		Accuracy measure for binary predicate extraction
			√						
				√					
					√			Inductive logic programming	
SYNDIKATE (2001)	√						Syntactic structure analysis, Anaphora resolution		F-measure for relation extraction Accuracy measure for concept extraction
		√					Use of semantic templates and domain knowledge	Inference engine	
			√						
				√					
OntoLearn (2002)	√					Relevance analysis	Part-of-speech tagging, Sentence parsing		F-measure for term extraction
		√					Concepts and glossary from WordNet		
			√				Hypernyms from WordNet		
CRCTOL (2005)	√	√				Relevance analysis	Part-of-speech tagging, Sentence parsing, Use of domain lexicon, Word sense disambiguation		F-measure for term and relation extraction
			√	√			Lexico-syntactic patterns, Syntactic structure analysis		
OntoGain (2010)	√	√					Part-of-speech tagging, Shallow parsing, Relevance analysis		Precision measure for concept extraction, hierarchy construction, and non-taxonomic relation extraction
			√			Aglomerative Clustering, Formal concept analysis			
				√		Association rule mining			

- —*Extract terms and form concepts.* The nouns in the arguments of the subcategorization frames extracted from the previous step are gathered to form basic classes based on the assumption *"head words occurring after the same, different prepositions (or with the same, different syntactic roles), and with the same, different verbs represent the same concept"* [Faure and Nedellec 1998a]. To illustrate, suppose that we have the nouns "ballpoint pen", "pencil and "fountain pen" occurring in different clauses as adjunct of the verb "to write" after the preposition "with". At the same time, these nouns are the direct object of the verb "to purchase". From the assumption, these nouns are thus considered as variants representing the same concept.
- —*Construct hierarchy.* The basic classes from the previous task are successively aggregated to form concepts of the ontology and reveal the taxonomic relations using clustering. Distance between all pairs of basic classes is computed, and two basic classes are only aggregated if the distance is less than the threshold set by the user. On the one hand, the distance between two classes containing the same words with the same frequencies have distance *0*. On the other hand, a pair of classes without a single common word have distance *1*. The clustering algorithm works bottom up and performs first-best using basic classes as input and builds the ontology level by level. User participation is required to validate each new cluster before it can be aggregated to a concept.

An evaluation of the term extraction technique was performed using the precision measure. The evaluation uses texts from the French journal *Le Monde* that have been manually filtered to ensure the presence of terrorist event descriptions. The results were evaluated by two domain experts who were not aware of the ontology building process using the following indicators: *OK* if extracted information is correct, *FALSE* if extracted information is incorrect, *NONE* if there were no extracted information, and *FALSE* for all other cases. Two precision values are computed, namely, *precision1* which is the ratio between *OK* and *FALSE*, and *precision2* which is the same as *precision1* by taking into consideration *NONE*. *Precision1* and *precision2* have values of 86% and 89%, respectively.

4.2. Text-to-Onto

Text-to-Onto [Cimiano and Staab 2005; Maedche and Staab 2000a, 2000b; Maedche and Volz 2001] is a semi-automated system that is part of an ontology management infrastructure called KAON.[4] KAON is a comprehensive tool suite for ontology creation and management. Text-to-Onto uses linguistics and statistics-based techniques to perform its ontology learning tasks as described next.

- —*Preprocess texts and extract terms.* Plain text extraction is performed to extract plain domain texts from semi-structured sources (i.e., HTML documents) and other formats (e.g., PDF documents). Abbreviation expansion is performed on the plain texts using rules and dictionaries to replace abbreviations and acronyms. Part-of-speech tagging and sentence parsing are performed on the preprocessed texts to produce syntactic structures and dependencies. Syntactic structure analysis is performed using weighted finite state transducers to identify important noun phrases as terms. These natural language processing tools are provided by a system called the Saarbruecken Message Extraction System (SMES) [Neumann et al. 1997].
- —*Form concepts.* Concepts from domain lexicon are required in order to assign new terms to predefined concepts. Unlike other approaches that employ general background knowledge, such as WordNet, the lexicon adopted by Text-to-Onto are domain-specific containing over 120,000 terms. Each term is associated with

[4] More information is available via `http://kaon.semanticweb.org/`. Last accessed May 25, 2009.

concepts available in a concept taxonomy. Other techniques for concept formations are also performed such as the use of cooccurrence analysis, but no additional information is provided.

—*Construct hierarchy.* Once the concepts have been formed, taxonomic relations are discovered by exploiting the hypernym from WordNet. Lexico-syntactic patterns are also employed to identify hypernymy relations in the texts. The authors refer to the hypernym as *oracle*, denoted by H. The projection $H(t)$ will return a set of tuples (x, y), where x is a hypernym for term t, and y is the number of times the algorithm has found evidence for it. Using cosine measure for similarity and the oracle, a bottom-up hierarchical clustering is carried out with a list T of n terms as input. When given two terms which are similar according to the cosine measure, the algorithm works by ordering them as subconcepts if one is a hypernym of the other. If the previous case does not apply, the most frequent common hypernym h is selected to create a new concept to accommodate both terms as siblings.

—*Discover non-taxonomic relations and label non-taxonomic relations.* For non-taxonomic relations extraction, association rules together with two user-defined thresholds (i.e., confidence, support) are employed to determine associations between concepts at the right level of abstraction. Typically, users start with low support and confidence to explore general relations and later increase the values to explore more specific relations. User participation is required to validate and label the non-taxonomic relations.

An evaluation of the relation discovery technique was performed using a measure called the *Generic Relations Learning Accuracy (RLA)*. Given a set of discovered relations D, precision is defined as $|D \cap R|/|D|$ and recall as $|D \cap R|/|R|$, where R is the non-taxonomic relations prepared by domain experts. RLA is a measure for capturing intuitive notions for relation matches, such as *utterly wrong*, *rather bad*, *near miss*, and *direct hit*. RLA is the averaged accuracy that the instances of discovered relations match against their best counterpart from manually curated gold standard. As the learning algorithm is controlled by support and confidence parameters, the evaluation is done by varying the support and the confidence values. When the support and confidence thresholds are set to 0, an RLA of 0.51 was obtained with $8,058$ relations. Both the number of relations and the recall decreases with growing support and confidence. Precision increases at first but drops when so few relations are discovered that almost none is a direct hit. The best RLA at 0.67 is achieved with a support at 0.04 and a confidence at 0.01.

4.3. TextStorm/Clouds

TextStorm/Clouds [Oliveira et al. 2001; Pereira et al. 2000] is a semi-automated ontology learning system that is part of an idea sharing and generation system called Dr. Divago [Pereira and Cardoso 1999]. The aim of this approach is to build and refine domain ontology for use in Dr. Divago for searching resources in a multidomain environment in order to generate musical pieces or drawings. TextStorm/Clouds uses logic and linguistics-based techniques to perform its ontology learning tasks as described next.

—*Preprocess texts and extract terms.* The part-of-speech information in WordNet is used to annotate the input text. Later, syntactic structure and dependency analysis is performed using an augmented grammar to extract syntactic structures in the form of *binary predicates*. The Prolog-like binary predicates represent relations between two terms. Two types of binary predicates are considered. The first type captures terms in the form of subject and object connected by a main verb. The second type captures the property of compound nouns in the form of modifiers. For example, the sentence “Zebra eat green grass” will result in two binary predicates, namely, `eat(Zebra, grass)` and `property(grass, green)`. When working with dependent sentences, finding the concepts may not be straightforward, and this approach performs anaphora resolution to resolve ambiguities. The anaphora resolution uses a history list of discourse entities generated from preceeding sentences [Allen 1995]. In the presence of an anaphora, the most recent entities are given higher priority.

—*Construct hierarchy, discover non-taxonomic relations, and label non-taxonomic relations.* Next, the binary predicates are employed to gradually aggregate terms and relations to an existing ontology with user participation. Hypernymy relations appear in binary predicates in the form of `is-a(X,Y)`, while `part-of(X,Y)` and `contain(X,Y)` provide good indicators for meronyms. Attribute value relations are obtainable from the predicates in the form of `property(X,Y)`. During the aggregation process, users may be required to introduce new predicates to connect certain terms and relations to the ontology. For example, in order to attach the predicate `is-a(predator, animal)` to an ontology with the root node `living_entity`, users would have to introduce `is-a(animal, living_entity)`.

—*Extract axioms.* The approach employs inductive logic programming to learn regularities by observing the recurrent concepts and relations in the predicates. For instance, the approach using the following extracted predicates

```
1: is-a(panther, carnivore)
2: eat(panther, zebra)
3: eat(panther, gazelle)
4: eat(zebra, grass)
5: is-a(zebra,herbivore)
6: eat(gazelle, grass)
7: is-a(gazelle,herbivore)
```

will arrive at the conclusions that

```
1: eat(A, zebra):- is-a(A, carnivore)
2: eat(A, grass):- is-a(A, herbivore).
```

These axioms describe relations between concepts in terms of its context (i.e., the set of neighbourhood connections that the arguments have).

Using the accuracy measure, the performance of the binary predicate extraction task was evaluated to determine whether the relations hold between the corresponding concepts. A total of 21 articles from the scientific domain were collected and analyzed by the system. Domain experts then determined the coherence of the predicates and its accuracy with respect to the corresponding input text. The authors concluded an average accuracy of 52%.

4.4. SYNDIKATE

SYNDIKATE [Hahn and Romacker 2001, 2000] is a stand-alone automated ontology learning system. SYNDIKATE uses only linguistics-based techniques to perform its ontology learning tasks as described next.

—*Extract terms.* Syntactic structure and dependency analysis is performed on the input text using a lexicalized dependency grammar to capture binary valency[5] constraints between a syntactic head (e.g., noun) and possible modifiers (e.g., determiners, adjectives). In order to establish a dependency relation between a head and a modifier, the term order, morpho-syntactic features compatibility, and semantic criteria have

[5] Valency refers to the capacity of a verb for taking a specific number and type of argument (noun phrase positions).

to be met. Anaphora resolution based on the centering model is included to handle pronouns.

—*Form concepts, construct hierarchy, discover non-taxonomic relations, and label non-taxonomic relations.* Using predefined semantic templates, each term in the syntactic dependency graph is associated with a concept in the domain knowledge and, at the same time, used to instantiate the text knowledge base. The text knowledge base is essentially an annotated representation of the input texts. For example, the term "hard disk" in the graph is associated with the concept *HARD_DISK* in domain knowledge, and at the same time, an instance called *HARD_DISK3* will be created in the text knowledge base. The approach then tries to find all relational links between conceptual correlates of two words in the subgraph if both grammatical and conceptual constraints are fulfilled. The linkage may either be constrained by dependency relations, by intervening lexical materials, or by conceptual compatibility between the concepts involved. In the case where unknown words occur, semantic interpretation of the dependency graph involving unknown lexical items in the text knowledge base is employed to derive concept hypothesis. The structural patterns of consistency, mutual justification, and analogy relative to the already available concept descriptions in the text knowledge base will be used as initial evidence to create linguistic and conceptual quality labels. An inference engine is then used to estimate the overall credibility of the concept hypotheses by taking into account the quality labels.

An evaluation using the precision, recall, and accuracy measures was conducted to assess the concepts and relations extracted by this system. The use of semantic interpretation to discover the relations between conceptual correlates yielded 57% recall and 97% precision, and 31% recall and 94% precision, for medicine and information technology texts, respectively. As for the formation of concepts, an accuracy of 87% was achieved. The authors also presented the performance of other aspects of the system. For example, sentence parsing in the system exhibits a linear time complexity, while a third-party parser runs in exponential time complexity. This behaviour was caused by the latter's ability to cope with ungrammatical input. The incompleteness of the system's parser results in a 10% loss of structural information, as compared to the complete third-party parser.

4.5. OntoLearn

OntoLearn [Missikoff et al. 2002; Navigli and Velardi 2002; Velardi et al. 2001, 2005] together with Consys (for ontology validation by experts) and SymOntoX (for updating and managing ontology by experts) are part of a project for developing an interoperable infrastructure for small and medium enterprises in the tourism sector under the Federated European Tourism Information System[6] (FETISH). OntoLearn uses linguistics and statistics-based techniques to perform its ontology learning tasks as described next.

—*Preprocess texts and extract terms.* Domain and general corpora are first processed using part-of-speech tagging and sentence parsing tools to produce syntactic structures, including noun phrases and prepositional phrases. For relevance analysis, the approach adopts two metrics known as *domain relevance (DR)* and *domain consensus (DC)*. Domain relevance measures the specificity of term t with respect to the target domain D_k through comparative analysis across a list of predefined domains

[6] More information is available via `http://sourceforge.net/projects/fetishproj/`. Last accessed May 25, 2009.

$D_1, \ldots, D_n$. The measure is defined as

$$DR(t, D_k) = \frac{P(t|D_k)}{\sum_{i=1\ldots n} P(t|D_i)},$$

where $P(t|D_k)$ and $P(t|D_i)$ are estimated as $\frac{f_{t,k}}{\sum_{t\in D_k} f_{t,k}}$ and $\frac{f_{t,i}}{\sum_{t\in D_i} f_{t,i}}$, respectively. $f_{t,k}$ and $f_{t,i}$ are the frequencies of term t in domain D_k and D_i, respectively. Domain consensus, on the other hand, is used to measure the appearance of a term in a single document, as compared to the overall occurrence in the target domain. The domain consensus of a term t in domain D_k is an entropy defined as

$$DC(t, D_k) = \sum_{d\in D_k} P(t|d) log \frac{1}{P(t|d)},$$

where $P(t|d)$ is the probability of encountering term t in document d of domain D_k.

—*Form concepts*. After the list of relevant terms has been identified, concepts and glossary from WordNet are employed for associating the terms to existing concepts and to provide definitions. The author named this process as *semantic interpretation*. If multi-word terms are involved, the approach evaluates all possible sense combinations by intersecting and weighting common semantic patterns in the glossary until it selects the best sense combinations.

—*Construct hierarchy*. Once semantic interpretation has been performed on the terms to form concepts, taxonomic relations are discovered using hypernyms from WordNet to organize the concepts into domain concept trees.

An evaluation of the term extraction technique was performed using the F-measure. A tourism corpus was manually constructed from the Web containing about 200,000 words. The evaluation was done by manually looking at 6,000 of the 14,383 candidate terms and marking all the terms judged as good domain terms and comparing the obtained list with the list of terms automatically filtered by the system. A precision of 85.42% and recall of 52.74% were achieved.

4.6. CRCTOL

CRCTOL [Jiang and Tan 2010], which stands for *concept-relation-concept tuple-based ontology learning*, is a system initially developed in 2005 at the National Technological University of Singapore for constructing ontologies from domain-specific documents. CRCTOL uses linguistics and statistics-based techniques to perform its ontology learning tasks as described next.

—*Preprocess texts*. A data importer is used to convert documents of different formats, such as PDF, XML, into plain texts. Stanford's part-of-speech tagger and the Berkeley Parser are used to tag words with part-of-speech and syntactic information.

—*Extract terms and form concepts*. Multi-word terms in the form of nouns and noun phrases are first extracted from the parsed texts using a set of predefined part-of-speech and syntactic tag-based rules. A manually built and maintained domain lexicon is used to identify terms which are specific to the domain. Articles and descriptive adjectives are then removed from the extracted terms. Next, a *domain relevance measure (DRM)* is used to weigh each term.

$$DRM(t) = \frac{tf(t)}{max(tf)} \times \frac{|\log\lambda(t)| - \min|\log\lambda|}{\max|\log\lambda| - \min|\log\lambda|} \times \frac{df(t)}{\max(df)},$$

where

$$\lambda(t) = \frac{\max_p p^{k_1}(1-p)^{n_1-k_1} p^{k_2}(1-p)^{n_2-k_2}}{\max_{p_1,p_2} p_1^{k_1}(1-p_1)^{n_1-k_1} p_2^{k_2}(1-p_2)^{n_2-k_2}},$$

where p_1 and p_2 are the probabilities of the occurrence of term t in the target d and the contrasting domain d'; k_1 and k_2 are the frequencies of t in d and d'; n_1 and n_2 are the total number of terms in d; and d', and p is the probability of t occurring in d and d'. $tf(t)$ and $df(t)$ are the term and document frequency of t from the TF-IDF measure, respectively. Terms with high DRM values are selected to form the initial concept list. A modified version of the LESK word disambiguation technique [Lesk 1986] is used for identifying the intended meaning of the extracted terms.

—*Construct hierarchy and discover non-taxonomic relations.* The authors use both lexico-syntactic patterns and head-modifier relations to identify the `is-a` relations between terms. For instance, hyponymy relations between the noun phrases in the form of $(parent, child) = (NP_0, NP_i)$ are extracted using the rule NP_0 such as NP_1 `(and|or)` $NP_2 \ldots$`(and|or)`$\ldots NP_n$. The patterns used in CRCTOL include the following.

1: NP$_0$ `(including|such as)` NP$_1$ NP$_2$ `...(and|or) ...`NP$_n$
2: NP$_1$ `is a kind of` NP$_0$,

where NP$_0$ is the hypernym of NP$_1$ to NP$_n$. As for non-taxonomic relations, the conventional approach of using rules to extract tuples in the form of `<noun1><verb><noun2>` is adopted. Verbs in the tuples are considered as lexical realizations of the semantic relations between the two concepts represented by `noun1` and `noun2`.

An evaluation in the domain of terrorism using the F-measure was conducted to assess the term and relation extraction components of CRCTOL against the Text-To-Onto system. Reports from the U.S. State Department were used as the domain corpus. The contrasting corpora were gathered from the TREC collection covering the commercial, computer, energy, and other general domains. The term extraction performance of CRCTOL was reported to be 99.5%, which is 1.9% higher than that of Text-To-Onto. CRCTOL achieved a 9.4% increase in F-score at 90.3%, as compared to Text-To-Onto, for simple sentences. CRCTOL's F-score for complex sentences stood at 68.6%, while Text-To-Onto reported only a 38.2% in performance.

4.7. OntoGain

The more recent OntoGain system [Drymonas et al. 2010] from the Technical University of Crete is designed for the unsupervised acquisition of ontologies from unstructured text. Similar to CRCTOL, OntoGain has been tested against Text2Onto, the successor of Text-To-Onto, in two different domains, namely, the medical and computer science domains. OntoGain uses linguistics and statistics-based techniques to perform its ontology learning tasks as described next.

—*Preprocess texts.* The OpenNLP suite of tools and the WordNet Java Library are first used for tokenization, lemmatization, part-of-speech tagging, and shallow parsing.

—*Extract terms and form concepts.* OntoGain implements the existing C/NC-value measure [Frantzi and Ananiadou 1997] for extracting compound or nested multiword terms. The C-value of a term t is given by

$$Cvalue(t) = \begin{cases} \log_2 |t| f_t & \text{if } |t| = g \\ \log_2 |t| \left(f_t - \frac{\sum_{l \in L_t} f_l}{|L_t|} \right) & \text{otherwise,} \end{cases} \quad (7)$$

where $|t|$ is the number of words that constitute t; L_t is the set of potential longer term candidates that contain t; g is the longest n-gram considered; and f_t is the frequency of occurrences of t in the corpus. The C-value measure is based upon the notion that a substring of a term candidate is a candidate itself, given that it demonstrates adequate independence from the longer version in which it appears [Wong et al. 2008b]. For instance, "E. coli food poisoning", "E. coli" and "food poisoning" are acceptable as valid complex term candidates. "E. coli food", however, is not. The NC-value, on the other hand, augments C-value by giving preference to terms that tend to cooccur within a specific context.

—*Construct hierarchy and discover non-taxonomic relations*. The authors implemented agglomerative clustering into OntoGain to build a hierarchy. As usual, each term is considered as a cluster initially, and with each step, clusters are merged based on a similarity measure. A lexical-based group average measure similar to the Dice-like coefficient that incorporates the constituents in multi-word terms is used.

$$sum(x, y) = \frac{|C(x_h) \cap C(y_h)|}{|C(x_h)| + |C(y_h)|} + \frac{|C(x) \cap C(y)|}{|C(x)| + |C(y)|}, \tag{8}$$

where x_h and y_h are the heads of term x and term y, respectively, and their set of constituents is denoted by C. *Formal concept analysis (FCA)* is also used to build hierarchies in OntoGain. A *formal contexts* matrix containing a set of formal objects, which are the extracted multi-word terms, and also containing attributes, which are the associated verbs identified during shallow parsing, is provided as input to the FCA algorithm. Similar to Text-To-Onto, association rule mining is used to discover non-taxonomic relations. The predictive apriori algorithm implementation on the Weka platform[7] is used for this purpose.

The OntoGain system was compared against Text2Onto in the domains of medicine (e.g., texts from MEDLINE) and computer science (e.g., scientific papers) with the help of domain experts. The evaluation reported precision values within the range of 86.67–89.7% depending on the domain for concept extraction. The construction of hierarchies using FCA recorded low precision values between 44.2–47.1%, while the performance of agglomerative clustering for this task is comparatively better in the range of 71.2–71.33%. Last, non-taxonomic relation extraction using association rule mining achieved precision values between 71.8–72.85%. No quantitative results were provided for the comparison between OntoGain and Text2Onto.

5. RECENT ADVANCES IN ONTOLOGY LEARNING TECHNIQUES

Since the publication of the five survey papers [Ding and Foo 2002; Gomez-Perez and Manzano-Macho 2003; Shamsfard and Barforoush 2003; Buitelaar et al. 2005; Zhou 2007], research activities within the ontology learning community have largely been focused on improving (1) term extraction and concept formation and (2) relation discovery techniques. The learning of ontologies (3) from social data and (4) across different languages has also been a topic of great research interest in the later part of the past decade. The recent progress in these four aspects will be discussed in the subsequent three sections.

5.1. Term Extraction and Concept Formation

Sclano and Velardi [2007] developed a technique called TermExtractor for identifying relevant terms in two steps. TermExtractor uses a sentence parser to parse texts and extract syntactic structures, such as noun compounds and `ADJ-N` and `N-PREP-N` sequences. The list of term candidates is then ranked and filtered using a combination of measures for realizing different evidence, namely, *domain pertinence (DP)*, *domain*

[7] http://www.cs.waikato.ac.nz/ml/weka/.

consensus (DC), *lexical cohesion (LC)*, and *structural relevance (SR)*. Wermter and Hahn [2005] incorporated a linguistic property of terms as evidence, namely, *limited paradigmatic modifiability*, into an algorithm for extracting terms. The property of paradigmatic modifiability is concerned with the extent to which the constituents of a multi-word term can be modified or substituted. The more we are able to substitute the constituents by other words, the less probable it is that the corresponding multi-word lexical unit is a term. Wong et al. [2009b] proposed a probabilistic framework for combining a variety of linguistically and heuristically motivated evidence to determine scores for ranking term candidates. In this framework, the characteristics that define a term are used to inspire the calculation of probabilities for ranking.

There is also an increase in interest in automatically constructing the text corpora required for term extraction using Web data. Agbago and Barriere [2005] proposed the use of richness estimators to assess the suitability of webpages provided by search engines for constructing corpora for use by terminologists. Baroni and Bernardini [2004] developed the BootCat technique for bootstrapping text corpora and terms using Web data and search engines. The technique requires as input a set of seed terms. The seeds are used to build a corpus using webpages suggested by search engines. New terms are then extracted from the initial corpus which in turn are used as seeds to build larger corpora. Realizing the shortcomings of the existing query-and-download approach, Wong et al. [2008a, 2011] developed a novel technique called SPARTAN which places emphasis on the analysis of the domain representativeness of websites for constructing virtual corpora. This technique also provides the means to extend the virtual corpora in a systematic way to construct specialized Web-derived corpora with high vocabulary coverage and specificity. The authors showed that SPARTAN is independent of the search engines used during corpus construction. The evaluations by the authors demonstrated that SPARTAN-based corpora achieved the best precision and recall in comparison to BootCat-derived corpora and the unconstrained querying of the Web for term recognition.

Zhang and Ciravegna [2011] offer an alternative view to concept formation as a task of named-entity recognition using the Web for background knowledge. The authors proposed a novel method that automatically creates domain-specific background knowledge by exploring Wikipedia for classifying terms into predefined ontological classes. The authors also demonstrated the potential use of this method for ontology population. Massey and Wong [2011], on the other hand, proposed a new topic extraction approach that allows 'meaning' to emerge naturally from the activation and decay of information in unstructured text retrieved from the Web. This approach may be used as an alternative method for discovering concepts using the unstructured texts in webpages as a source of knowledge. The authors discussed the results from several initial experiments comparing the use of WordNet versus webpages from Yahoo! search on the Reuters-21578 corpus to illustrate the power of this new approach. Other techniques for latent topic extraction, such as *latent dirichlet allocation (LDA)*, have also been used to discover concepts in ontology learning [Yeh and Yang 2008].

5.2. Relation Discovery

Specia and Motta [2006] presented a pipeline of existing tools for extracting semantic relations between pairs of entities from texts. The approach uses the tokenizer, part-of-speech tagger, and verb phrase chunker from GATE [Cunningham et al. 2002] together with Minipar [Lin 1998] to provide the annotations required for extracting linguistic triples. Terms from the triples are mapped to their corresponding concepts using a domain ontology and a named-entity recognition system. Any ambiguity in the relations between a pair of entities is resolved using SenseLearner [Mihalcea and Csomai 2005]. Relations are detected using a collection of predefined patterns as well as the existing knowledge in a domain ontology and lexical databases. Extracted entities that exist in the knowledge base are semantically annotated with their properties. Ciaramita et al. [2005] employ syntactic dependencies as potential relations. The dependency paths are treated as bi-grams and scored with statistical measures of correlation. At the same time, the arguments of the relations can be generalized to obtain abstract concepts using algorithms for *selectional restrictions learning* [Ribas 1995]. Snow et al. [2005, 2006] also presented an approach that employs the dependency paths extracted from parse trees. The approach receives trainings using sets of text containing known hypernym pairs. The approach then automatically discovers useful dependency paths that can be applied to new corpora for identifying new hypernym.

The trend of using Web data to improve the discovery of semantic relations is also on the rise. Sombatsrisomboon et al. [2003] proposed a simple three-step technique for discovering taxonomic relations (i.e., hypernym/hyponym) between pairs of terms using search engines. Search engine queries are first constructed using the term pairs and patterns, such as `X is a/an Y`. The webpages provided by search engines are then gathered to create a small corpus. Sentence parsing and syntactic structure analysis is performed on the corpus to discover taxonomic relations between the terms. Such use of patterns, and Web data redundancy can also be extended to discover non-taxonomic relations. Sanchez and Moreno [2008] proposed methods for discovering non-taxonomic relations using Web data. The authors developed a technique for learning domain patterns using domain-relevant verb phrases extracted from webpages provided by search engines. These domain patterns are then used to extract and label non-taxonomic relations using linguistic and statistical analysis. Etzioni et al. [2008], on the other hand, developed TextRunner to extract information across different domains from the Web. The entities and relationships extracted using TextRunner are useful for bootstrapping the construction ontologies. TextRunner operates in a two-phase architecture. The first phase uses a conditional random field-based model to label the constituents in the input strings as either entities or relationships. An extractor is then used in the second phase to extract triples to capture the relationships between entities.

We have noticed an increasing interest in the use of structured Web data, such as Wikipedia, for relation acquisition. Pei et al. [2008] proposed an approach for constructing ontologies using Wikipedia. The approach uses a two-step technique, namely, name mapping and logic-based mapping to deduce the type of relations between concepts in Wikipedia. Similarly, Liu et al. [2008] developed a technique called Catriple for automatically extracting triples using Wikipedia's categorical system. The approach focuses on category pairs containing both explicit property and explicit value (e.g., "Category: Songs by artist"-"Category: The Beatles songs", where "artist is property and "The Beatles" is value), and category pairs containing explicit value but implicit property (e.g. "Category: Rock songs"-"Category: British rock songs" where "British" is a value with no property). Sentence parsers and syntactic rules are used to extract the explicit properties and values from the category names. Weber and Buitelaar [2006] proposed a system called Information System for Ontology Learning and Domain Exploration (ISOLDE) for deriving domain ontologies using manually curated text corpora, a general-purpose named-entity tagger, and structured data on the Web (i.e., Wikipedia, Wiktionary, and a German online dictionary known as DWDS) to derive a domain ontology. Wong et al. [2009a] proposed a hybrid approach based on techniques in lexical simplification, word disambiguation, and association inference for acquiring coarse-grained relations between potentially ambiguous and composite terms using only Wikipedia and search engine page count. Mintz et al. [2009] uses Freebase instead of the typical WordNet as a lookup dictionary for discovering relations between pairs of entities.

5.3. Ontology Learning from Social Data and Across Different Languages

In addition to improvements over the existing techniques for extracting concepts and relations, there is also an increase in interest in the social aspect of ontology learning. For instance, Tang et al. [2009] investigated the problem of ontology learning from user-defined tags on Web 2.0 portals, also known as *folksonomies*. The authors proposed an approached based on a probabilistic topic model to represent the tags and their annotated documents. Four divergence measures were also defined to characterize the relations between tags. Data from citeulike.com and imdb.com were used in their experiments to show that ontological hierarchy can be effectively learned from social tags using the proposed approach. Kotis and Papasalouros [2011], on the other hand, discussed more broadly the requirements for automatically learning ontologies from social data on the Web, such as blogs, wikis, and folksonomies. The authors presented two techniques for automatically learning ontologies of social concepts and relations from query logs and Web 2.0 question/answer applications such as Yahoo! Answer. The authors evaluated the ontology learning technique from query logs using Yahoo! and Google query datasets. The authors also discussed the importance of modeling trust for specifying the degree of confidence that agents, both software and human, may have on the conceptualizations derived from social content. The role of users in ontology creation becomes much more obvious when we examine the tripartite model of ontologies proposed by Mika [2007]. This abstract model of semantic-social networks, which the author referred to as the *actor concept instance model*, is built upon the realization that the meaning associated with concepts and relations is necessarily dependent on a community of actors (i.e., emergent semantics). Weichselbraun et al. [2010] described an approach that complements corpus-based ontology learning with tags derived from Web 2.0 services, such as social networking platforms and microblogging services. These tags provide an external view of the domain and can be incorporated as external knowledge into the ontology learning process.

Besides the social dimension of ontology creation, ontology learning from multilingual text is also gaining popularity. Hjelm and Volk [Hjelm and Volk 2011; Hjelm 2009] discussed ways to automatically construct ontologies by exploiting cross-language information from parallel corpora. In particular, the authors presented a framework that provides a setting in which cross-language data can be integrated and quantified for cross-language ontology learning. The authors employed resources, such as the JRC-ACQUIS Multilingual Parallel Corpus and the Eurovoc multilingual thesaurus for their experiments. The authors concluded that the combining of information from different languages can indeed improve the results of ontology learning. Lu et al. [2011] focused specifically on the mining of parallel sentences and parallel technical terms from comparable Chinese-English patent texts which contain both equivalent sentences as well as noise. The authors touched on the potential use of the extracted parallel sentences and technical terms for further acquisition of terms and relations, translation of monolingual ontologies, as well as other cross-lingual information access applications. In particular, the authors discussed the potentials and challenges of using linguistically diverse Web data to address the problem of mining the same knowledge across different languages.

6. CURRENT TRENDS AND FUTURE RESEARCH DIRECTIONS

To summarize, we began this survey with an overview of ontologies and ontology learning from text. In particular, we introduced a unified way of looking at the types of output, tasks, techniques, and resources in ontology learning as well as the associations between these different dimensions in Figure 2. We summarized several widely used evaluation methods in the ontology learning community. The differences between a formal and a lightweight ontology were also explained. Finally, we reviewed seven prominent ontology learning systems as well as recent advances in the field. A summary of the systems reviewed is provided in Table I.

In this section, we bring this survey to a close by summarizing the progress and trends that the ontology learning community has witnessed over the past ten years. We then look at several open issues that will likely define the future research directions of the community.

6.1. Trends in Ontology Learning Techniques

From the review of recent techniques in Section 5, we are able to observe that current research efforts are either in the stages of enhancing existing term recognition techniques or moving to the more advanced phase of relation discovery. It remains a trend that research into axiom learning is scarce. Let us first look at a summary of the recent research focus in ontology learning in terms of term and concept extraction and relation extraction.

First, the measures for scoring and extracting terms from texts have more or less stabilized, with performance generally above 90% in F-score. The current state of the art is based mainly on statistical semantics and paradigmatic and syntagmatic relations, that is to say, we determine the relevance of terms through observations in very large samples and through the way the constituents of a term are put together. To further improve the performance of term extraction, we are seeing a rise in interest for constructing very large text corpora from the Web. The techniques for constructing text corpora vary from the simple query-and-download approach to more complicated ones that require the analysis of webpage content. Term extraction techniques, especially those based on statistical semantics, benefit greatly from the work in this area. Considering that content-bearing domain terms are rare in texts, it has been shown that larger samples (i.e., text corpora) will improve the performance of such techniques. In addition to the typical clustering algorithms, techniques from named-entity recognition and topic extraction are also increasingly being used for generalizing terms to form concepts. The typical process of extracting terms and later forming concepts into hierarchies can be collapsed into a single task of parsing text to annotate noun phrases with predefined categories (i.e., concepts) such as "person" and "organisation". In this way, a term or named entity is immediately associated with a concept or category through the `is-a` relation.

Second, as for taxonomic and non-taxonomic relation discovery, we are witnessing the increasing application of lexico-syntactic patterns, association rule mining, and rules based on syntactic dependencies on very large datasets from the Web. Initially applied on small and restricted datasets, the redundancy of Web data has allowed this group of techniques that rely on repetitions and regularities to be revived and flourish. Since the second part of the decade, the preferred source of data for this class of techniques is Web search results. The accessibility of Web search engines has promoted the increased use of unstructured data for these tasks. For extracting domain-specific relations, specialized webpage collections can be built using the corpus construction techniques previously discussed for term extraction. Another type of resource on the Web that is fast becoming a necessary part of emerging work for discovering relations is (semi-)structured Web data, such as Wikipedia and Freebase. Figure 3, which shows the publications in the past ten years describing Wikipedia and relation extraction, demonstrates this trend. While both unstructured and (semi-)structured resources may appear to be the panacea for discovering relations between terms or concepts, many questions remain unaddressed.

Based on the trends previously discussed, we do not observe any particular preference towards either statistics- or linguistics-based techniques. The increasing popularity of

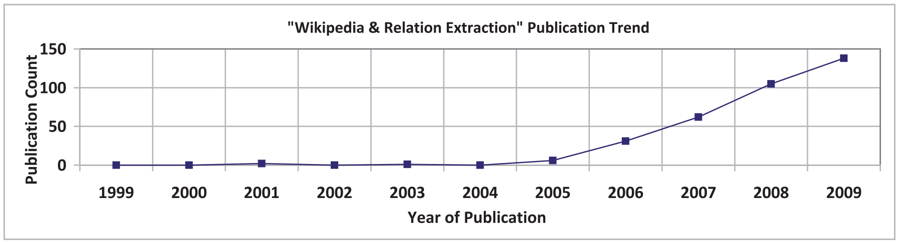


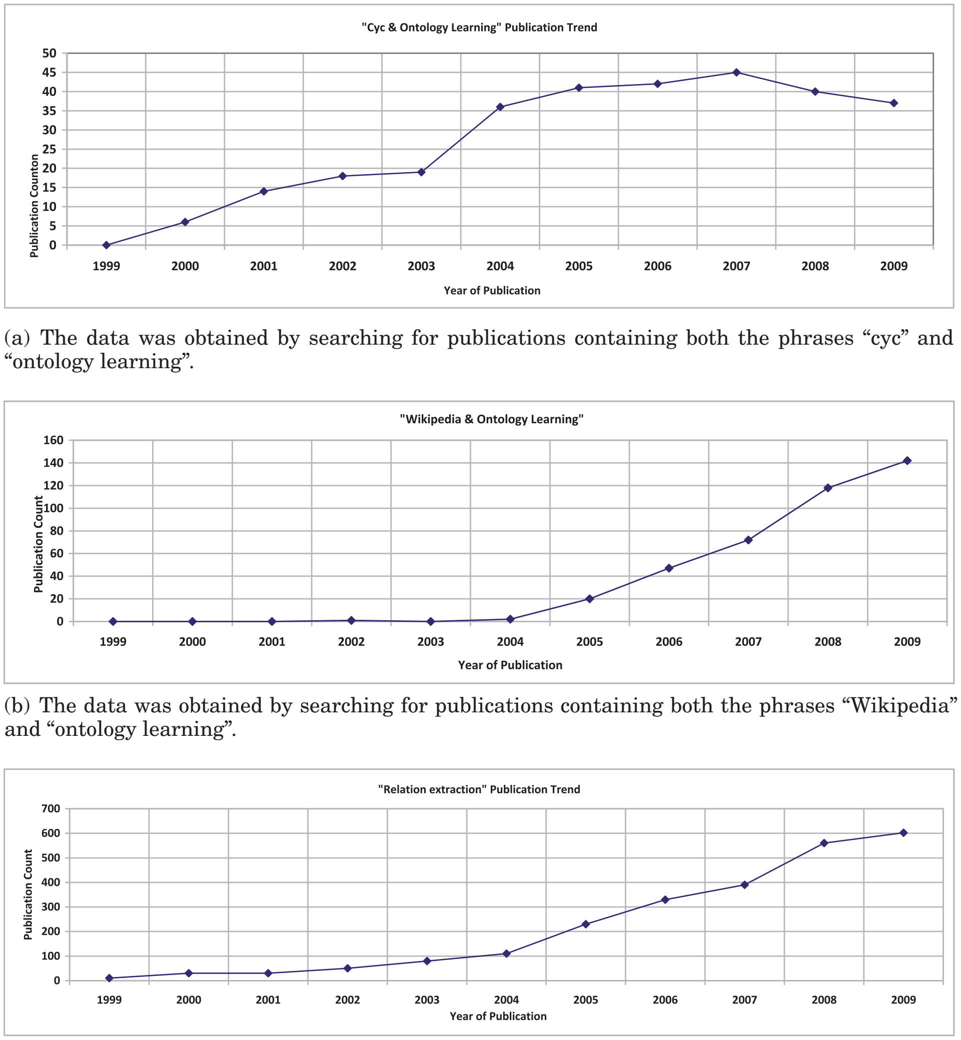
Fig. 3. The publication trend of Wikipedia and relation extraction. The graph is plotted using data from Google Scholar. The data was obtained by searching for publications containing both phrases "Wikipedia" and "relation extraction".

Web resources for complementing or even replaceing expert-crafted semantic lexicons or annotated corpora is, however, visible. This trend does not come as a surprise, as we gradually move into the learning of ontologies with minimal human intervention across different domains. Naturally, techniques that easily benefited from larger datasets received instant attention. In this sense, we can say that some techniques are indeed witnessing an increase in preference. We, however, do foresee a potential rise in relevance of logic-based techniques to the process of learning ontologies. In the final section, we will discuss this together with the reasons behind the need for more research in relation to the use of Web data for ontology learning.

6.2. Progress from a System Point of View

We may be tempted to jump to the conclusion that progress in ontology learning is nonexistent the very moment we look at the fact that five out of the seven systems reviewed in Section 4 and summarized in Table 1 are between eight to ten years old. This conclusion becomes more acceptable when we consider that the same pattern was observed in all the five previous surveys. To provide a more balanced view of the actual state of progress, we included a review of the recent advances in ontology learning techniques, as summarized in Section 6.1. The review in Section 4 and the summary in Section 6.1 suggest that despite the slow progress from a system point of view, we are able to observe considerable advances in the higher-layer tasks, as in the well as techniques that rely less on human involvement. This finding in fact correlates with the publication data shown in Figure 4. Figure 4(a), for instance, shows that publications citing the expert-crafted knowledge Cyc is slowing down. At the same time, the mention of Wikipedia in ontology learning publications has increased drastically since its conception in 2001, as shown in Figure 4(b). Similarly, Figure 4(c) shows that work on the higher-layer output, namely, relation extraction, has been on the increase. To further validate this pattern, we look back at the seven systems reviewed in Section 4. OntoLearn and ASIUM, which were conceived earlier in the decade, only support the construction of hierarchies and other lower-level tasks. CRCTOL and OntoGain, which were developed in the second part of the decade, are able to extract non-taxonomic relations. Even though the much older TextStorm/Clouds and SYNDIKATE are able to extract non-taxonomic relations and even axioms, these two systems require tremendous manual efforts in order to craft the required domain knowledge and define missing knowledge.

We can actually trace the catalyst of such a pattern of progress to one source, namely, the initial excitement and high hopes for ontology learning systems at the turn of the millenium. This conjecture is reasonable considering the high number of systems reported in the literature during the first half of the decade. Many of the earlier systems, as demonstrated by our review in Section 4, are proof of concepts for demonstrating to



(a) The data was obtained by searching for publications containing both the phrases "cyc" and "ontology learning".

(b) The data was obtained by searching for publications containing both the phrases "Wikipedia" and "ontology learning".

(c) The data was obtained by searching for publications containing the "relation extraction".

Fig. 4. The publication trend of the various aspects of ontology learning. The graphs are plotted using data from Google Scholar.

the research community and potential industry partners what ontology learning has to offer. Many of them are essentially mashups comprising hand-crafted knowledge and existing tools and techniques from advanced related areas. While these systems may be able to address the requirements of constructing small toy ontologies, time eventually reveals the need for researchers to return to the basics and address more fundamental issues, such as the knowledge acquisition bottleneck. This explains the reduction in the number on complete ontology learning systems reported in the literature during the second part of the decade. The fact remains that we have only started in the past few years to improve on the task of relation extraction for ontology learning, especially through the use of Web resources. For this reason, it is quite surprising initially to note that earlier systems, such as TextStorm/Clouds, are able to extract axioms. We believe that axiom learning remains a task to be addressed in the near future, as evidenced by the lack of mention in Section 6.1.

In short, we are able to speculate that despite the slow pace of development from a system point of view, there is indeed progress in the extraction of outputs in the higher layers with less and less human intervention, as summarized in Section 6.1. Overall, it is safe to conclude that the automatic construction of full-fledged ontologies from text across different domains is currently beyond the reach of conventional systems based on our reviews. This conclusion, which was also noted in the previous five surveys, does not come as a surprise considering that an ontology is after all a shared conceptualization of a domain. The involvement of consensus and high-level abstraction requires human cognitive processing. This makes the process of fully automating ontology learning impossible. Moreover, with the need for axioms in formal ontologies, coupled with our current inability to efficiently learn axioms, there is still plenty of work required to produce a system that can truly claim to learn full-fledged ontologies. We will nevertheless discuss in the next section a potential way of addressing the consensus aspect of ontology learning.

6.3. Outlook

The intertwining of the Web with ontology learning is a natural progression for many reasons. The ability to harvest consensus (considering that ontologies are shared conceptualizations) and accessibility to very large samples required by many learning techniques are amongst the reasons. In addition to the already existing problems in ontology learning, the growing use of Web data will introduce new challenges. At the moment, research involving the use of Web data for addressing the bottleneck of manual knowledge crafting has already begun. For instance, we are already seeing the marrying of Web data with term, concept, and relation extraction techniques that can easily benefit from larger datasets. For all we know, the Web may very well be the key ingredient in constructing ontologies with minimal human intervention required for cross-language and cross-domain applications and, eventually, the Semantic Web. When this happens, the role of formal ontology language will become much more significant, and heavyweight ontologies will take the center stage. We close this survey by looking at some of the present and future research problems in the area in this section.

First, we foresee that more and more research efforts will be dedicated to creating new or adapting existing techniques to work with the noise, richness, diversity, and scale of Web data. In regard to noise, there is currently little mention of data cleanliness during ontology learning. As the use of Web data becomes more common, integrated techniques for addressing spelling errors, abbreviations, grammatical errors, word variants, and so on in texts are turning into a necessity. For instance, looking for a more representative word count on the Web for “endeavour” will require consideration for its variants (e.g., “endeavor”) and spelling errors (e.g., “endevour”). Moreover, the issues of authority and validity in Web data sources must also be investigated. Otherwise, relations frequently occurring on the Web, such as `<Vladimir Putin><is-a><president of Germany>`, will end up in the knowledge base. We predict that social data from the Web (e.g., collaborative tagging) will play an increasingly important role in addressing the authority and validity aspects of ontology learning. Probabilities and ranking based on wisdom of the masses is one way to assign trust to concepts and relations acquired from Web sources.

Second, the richness of Web data in terms of (semi-)structured, collaboratively maintained resources, such as Wikipedia, is increasingly being used to improve higher-layer tasks, such as concept formation and relation discovery. We observed from the literature, the current mushrooming of techniques for finding semantic relations using the categorical structure of Wikipedia. These techniques are mostly focused on hierarchical relations and often leave out the details on how to cope with concepts that do not appear in Wikipedia. We foresee that more effort will be dedicated to studying and exploiting associative relations on Wikipedia (e.g., links under the "See also" section) for ontology learning. We have already noticed work on identifying coarse-grained unlabeled associative relations from Wikipedia and the adaptive matching of terms to Wikipedia topics where exact matches are not available. We will definitely see more work going along this direction. An example would be the use of the coarse-grained associative relations as seeds together with triples extracted from Web search results for bootstrapping the discovery of more detailed semantic relations. The verbs from the triples could then be used to label the relations. Unless improvements are made in these tasks, many of the current elaborate and expert-crafted ontologies, such as the Gene Ontology, cannot be replicated using ontology, learning from text systems.

Third, the diversity of Web data has also contributed to the rise of cross-language ontology learning in the past few years. As more communities of different cultural and linguistic backgrounds contribute to the Web, the availability of textual resources required for ontology learning across different languages will improve. The potential growth of cross-language research in the future signals the need to gradually move ontologies away from language dependency. Considering that formal ontologies are shared conceptualizations and should not contain lexical knowledge [Hjelm and Volk 2011], apples should not be represented lexically as "apple" in an ontology so that the overall `fruit` ontology can be applicable to other languages. For this to happen, we need more research into mechanisms for encoding and representing ontological entities as low-level constructs and for mapping these constructs into natural language symbols to facilitate human interpretation.

Fourth, the ability to cope with the scale of Web data required for ontology learning is also another concern. The efficiency and robustness in processing an exponentially growing volume of text will likely receive increasing attention. The issues that researchers will look at extend beyond mere storage space or other hardware considerations. Some of the topics of potential interest include the ease of analyzing petabyte collections for corpus statistics, the ability to commit, resume, and rollback the learning process in the event of errors or interruptions, and the efficiency of techniques for the various tasks of ontology learning from Web-scale data (e.g., large-scale sentence parsing). The latter topic is of particular interest considering that many of the current ontology learning systems employ readily available off-the-shelf tools or incorporate techniques designed for small datasets or without efficiency in mind. In particular, systems that are the result of putting together existing tools may not be streamlined and hence may suffer in performance when faced with Web-scale text analysis.

Fifth, we speculate that the related area of ontology mapping, also known as ontology alignment, will become more pertinent as the availability of ontologies increases. The availability of multiple and potentially conflicting or complementing ontologies will call for better means to determine correspondences between concepts and even relations [deBruijn et al. 2006]. A gradual rise in interest in ontology mapping is obvious as we look at the publication trend shown in Figure 5. The data for this graph are obtained by searching for publications containing the phrase "ontology mapping" or "ontology alignment" on Google Scholar. The graphs in Figures 4 and 5 may not be representative of the actual publication trends. They, however, do offer testable predictions of the current state of (as well as) future research interests. In addition, we are predicting a rise in focus on logic-based techniques in ontology learning, as our techniques for the lower layers (i.e., term, concept, relation) mature and our systems become more comprehensive (i.e., inclusion of higher-layer outputs), reaching towards the learning of full-fledged ontologies.

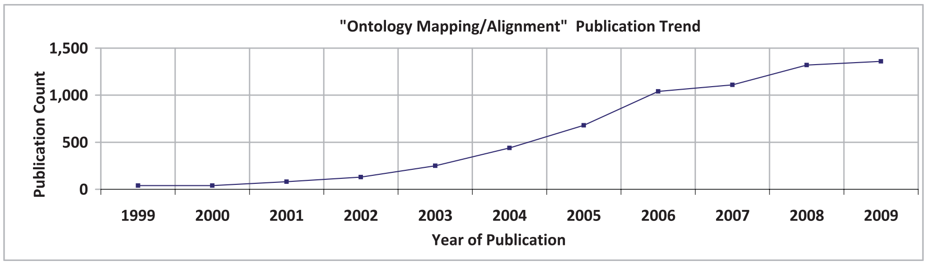


Fig. 5. The publication trend of ontology mapping. The graph is plotted using data from Google Scholar. The data was obtained by searching for publications containing the phrase "ontology mapping" or "ontology alignment".

Last, it remains a fact that the majority of the ontologies out there at the moment are lightweight. To be able to semi-automatically learn formal ontologies, we have to improve on our current axiom learning techniques as well as to find ways of incorporating the consensus aspect into the learning process, amongst others. As formal ontologies take the center stage, we foresee an increase in concern regarding the extensibility of existing lightweight ontologies to full-fledged ones.

All in all, there are several key issues that will likely define the research directions in this area in the near future, namely, (1) the issue of noise, authority, and validity in Web data for ontology learning; (2) the integration of social data into the learning process to incorporate consensus into ontology building; (3) the design of new techniques for exploiting the structural richness of collaboratively maintained Web data; (4) the representation of ontological entities as language-independent constructs; (5) the applicability of existing techniques for learning ontologies for different writing systems (e.g., alphabetic, logographic); (6) the efficiency and robustness of existing techniques for Web-scale ontology learning; (7) the increasing role of ontology mapping as more ontologies become available; and (8) the extensibility of existing lightweight ontologies to formal ones. Key phrases, such as Web-scale, open, consensus, social, formal, and cross-language ontology learning or ontologies, are all buzzwords that we will encounter very often in the future.

REFERENCES

Abou-Assali, A., Lenne, D., and Debray, B. 2007. KOMIS: An ontology-based knowledge management system for industrial safety. In *Proceedings of the 18th International Workshop on Database and Expert Systems Application*.

Agbago, A. and Barriere, C. 2005. Corpus construction for terminology. In *Proceedings of the Corpus Linguistics Conference*.

Agustini, A., Gamallo, P., and Lopes, G. 2001. Selection restrictions acquisition for parsing and information retrieval improvement. In *Proceedings of the 14th International Conference on Applications of Prolog*.

Allen, J. 1995. *Natural Language Understanding.* Benjamin Cummings, San Francisco, CA.

Baker, C., Kanagasabai, R., Ang, W., Veeramani, A., Low, H., and Wenk, M. 2007. Towards ontology-driven navigation of the lipid bibliosphere. In *Proceedings of the 6th International Conference on Bioinformatics (InCoB)*.

Baroni, M. and Bernardini, S. 2004. Bootcat: Bootstrapping corpora and terms from the Web. In *Proceedings of the 4th Language Resources and Evaluation Conference (LREC)*.

Basili, R., Moschitti, A., Pazienza, M., and Zanzotto, F. 2001. A contrastive approach to term extraction. In *Proceedings of the 4th Terminology and Artificial Intelligence Conference (TIA)*.

Berners-Lee, T., Hendler, J., and Lassila, O. 2001. The semantic Web. `http:www.scientificamerican.comarticle.cfm?id=the-semantic-web`. Last accessed 5/09.

Bird, S., Klein, E., Loper, E., and Baldridge, J. 2008. Multidisciplinary instruction with the natural language toolkit. In *Proceedings of the 3rd ACL Workshop on Issues in Teaching Computational Linguistics*.

BRANK, J., GROBELNIK, M., AND MLADENIC, D. 2005. A survey of ontology evaluation techniques. In *Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD)*.

BREWSTER, C., ALANI, H., DASMAHAPATRA, S., AND WILKS, Y. 2004. Data driven ontology evaluation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

BREWSTER, C., CIRAVEGNA, F., AND WILKS, Y. 2002. User-centred ontology learning for knowledge management. In *Proceedings of the 6th International Conference on Applications of Natural Language to Information Systems*.

BRILL, E. 1992. A simple rule-based part of speech tagger. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*.

BUDANITSKY, A. 1999. Lexical semantic relatedness and its application in natural language processing. Tech. rep. CSRG-390, Computer Systems Research Group, University of Toronto.

BUITELAAR, P., CIMIANO, P., AND MAGNINI, B. 2005. Ontology learning from text: An overview. In *Ontology Learning from Text: Methods, Evaluation and Applications*, P. Buitelaar, P. Cimmiano, and B. Magnini, Eds. IOS Press, Amsterdam.

CASTELLS, P., FERNANDEZ, M., AND VALLET, D. 2007. An adaptation of the vector-space model for ontology-based information retrieval. *IEEE Trans. Knowl. Data Eng. 19,* 2, 261–272.

CHO, J., HAN, S., AND KIM, H. 2006. Meta-ontology for automated information integration of parts libraries. *Comput.-Aided Des. 38,* 7, 713–725.

CHURCH, K. AND HANKS, P. 1990. Word association norms, mutual information, and lexicography. *Comput. Ling. 16,* 1, 22–29.

CIARAMITA, M., GANGEMI, A., RATSCH, E., SARIC, J., AND ROJAS, I. 2005. Unsupervised learning of semantic relations between concepts of a molecular biology ontology. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI)*.

CIMIANO, P., PIVK, A., SCHMIDT-THIEME, L., AND STAAB, S. 2004. Learning taxonomic relations from heterogeneous evidence. In *Proceedings of the ECAI Workshop on Ontology Learning and Population*.

CIMIANO, P. AND STAAB, S. 2005. Learning concept hierarchies from text with a guided agglomerative clustering algorithm. In *Proceedings of the Workshop on Learning and Extending Lexical Ontologies with Machine Learning Methods*.

CONSTANT, P. 1995. L'analyseur linguistique sylex. In *Proceedings of the 5eme Ecole d'ete du CNET*.

CROFT, B. AND PONTE, J. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st International Conference on Research and Development in Information Retrieval.*

CULLEN, J. AND BRYMAN, A. 1988. The knowledge acquisition bottleneck: Time for reassessment? *Expert Syst. 5,* 3, 216–225.

CUNNINGHAM, H., MAYNARD, D., BONTCHEVA, K., AND TABLAN, V. 2002. GATE: An architecture for development of robust hlt applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL)*.

DAVIES, J., FENSEL, D., AND VANHARMELEN, F. 2003. *Towards the Semantic Web: Ontology-driven Knowledge Management.* Wiley, Chichester.

DEBRUIJN, J., EHRIG, M., FEIER, C., MARTNS-RECUERDA, F., SCHARFFE, F., AND WEITEN, M. 2006. Ontology mediation, merging, and aligning. In *Semantic Web Technologies: Trends and Research in Ontology-based Systems*, J. Davies, R. Studer, and P. Warren, Eds. John Wiley & Sons, Hoboken, NJ.

DELLSCHAFT, K. AND STAAB, S. 2006. On how to perform a gold standard based evaluation of ontology learning. In *Proceedings of the 5th International Semantic Web Conference (ISWC)*.

DELLSCHAFT, K. AND STAAB, S. 2008. Strategies for the evaluation of ontology learning. In *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, P. Buitelaar and P. Cimiano, Eds. IOS Press, Amsterdam.

DING, Y. AND FOO, S. 2002. Ontology research and development: Part 1. *J. Inf. Sci. 28,* 2, 123–136.

DRYMONAS, E., ZERVANOU, K., AND PETRAKIS, E. 2010. Unsupervised ontology acquisition from plain texts: The OntoGain system. In *Proceedings of the 15th International Conference on Applications of Natural Language to Information Systems (NLDB)*.

ETZIONI, O., BANKO, M., SODERLAND, S., AND WELD, D. 2008. Open information extraction from the Web. *Commun. ACM 51,* 12, 68–74.

FAURE, D. AND NEDELLEC, C. 1998a. ASIUM: Learning subcategorization frames and restrictions of selection. In *Proceedings of the 10th Conference on Machine Learning (ECML)*.

FAURE, D. AND NEDELLEC, C. 1998b. A corpus-based conceptual clustering method for verb frames and ontology acquisition. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC)*.

FAURE, D. AND NEDELLEC, C. 1999. Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system ASIUM. In *Proceedings of the 11th European Workshop on Knowledge Acquisition, Modeling and Management (EKAW)*.

FAURE, D. AND POIBEAU, T. 2000. First experiments of using semantic knowledge learned by ASIUM for information extraction task using Intex. In *Proceedings of the 1st Workshop on Ontology Learning*.

FLUIT, C., SABOU, M., AND VANHARMELEN, F. 2003. Supporting user tasks through visualisation of lightweight ontologies. In *Handbook on Ontologies in Information Systems*, S. Staab and R. Studer, Eds. Springer-Verlag, Berlin.

FORTUNA, B., MLADENIC, D., AND GROBELNIK, M. 2005. Semi-automatic construction of topic ontology. In *Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD)*.

FOTZO, H. AND GALLINARI, P. 2004. Learning generalizationspecialization relations between concepts—application for automatically building thematic document hierarchies. In *Proceedings of the 7th International Conference on Computer-Assisted Information Retrieval (RIAO)*.

FRANTZI, K. AND ANANIADOU, S. 1997. Automatic term recognition using contextual cues. In *Proceedings of the IJCAI Workshop on Multilinguality in Software Industry: The AI Contribution*.

FUHR, N. 1992. Probabilistic models in information retrieval. *Comput. J. 35,* 3, 243–255.

FURST, F. AND TRICHET, F. 2006. Heavyweight ontology engineering. In *Proceedings of the International Conference on Ontologies, Databases, and Applications of Semantics (ODBASE)*.

GAMALLO, P., AGUSTINI, A., AND LOPES, G. 2003. Learning subcategorisation information to model a grammar with co-restrictions. *Traitement Automatic de la Langue 44,* 1, 93–117.

GAMALLO, P., GONZALEZ, M., AGUSTINI, A., LOPES, G., AND DELIMA, V. 2002. Mapping syntactic dependencies onto semantic relations. In *Proceedings of the ECAI Workshop on Machine Learning and Natural Language Processing for Ontology Engineering*.

GIUNCHIGLIA, F. AND ZAIHRAYEU, I. 2007. Lightweight ontologies. Tech. rep. DIT-07-071, University of Trento.

GOMEZ-PEREZ, A. AND MANZANO-MACHO, D. 2003. A survey of ontology learning methods and techniques. Deliverable 1.5, OntoWeb Consortium.

GRUBER, T. 1993. A translation approach to portable ontology specifications. *Knowl. Acquisit. 5,* 2, 199–220.

HAHN, U. AND ROMACKER, M. 2000. Content management in the SyndiKate system: How technical documents are automatically transformed to text knowledge bases. *Data Knowl. Eng. 35,* 1, 137–159.

HAHN, U. AND ROMACKER, M. 2001. The SyndiKate text knowledge base generator. In *Proceedings of the 1st International Conference on Human Language Technology Research*.

HATALA, M., SIADATY, M., GASEVIC, D., JOVANOVIC, J., AND TORNIAI, C. 2009. Utility of ontology extraction tools in the hands of educators. In *Proceedings of the IEEE International Conference on Semantic Computing (ICSC)*.

HEARST, M. 1998. Automated discovery of WordNet relations. In *WordNet: An Electronic Lexical Database and Some of its Applications*, C. Fellbaum, Ed. MIT Press, Cambridge, MA.

HIPPISLEY, A., CHENG, D., AND AHMAD, K. 2005. The head-modifier principle and multilingual term extraction. *Natural Lang. Eng. 11,* 2, 129–157.

HJELM, H. 2009. Cross-language ontology learning: Incorporating and exploiting cross-language data in the ontology learning process. Ph.D. dissertation, Stockholm University.

HJELM, H. AND VOLK, M. 2011. Cross-language ontology learning. In *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances*, W. Wong, W. Liu, and M. Bennamoun, Eds. IGI Global, Hershey, PA.

HWANG, C. 1999. Incompletely and imprecisely speaking: Using dynamic ontologies for representing and retrieving information. In *Proceedings of the 6th International Workshop on Knowledge Representation meets Databases (KRDB)*.

HYVONEN, E., STYRMAN, A., AND SAARELA, S. 2003. Ontology-based image retrieval. In *Proceedings of the 12th International World Wide Web Conference*.

JIANG, T., TAN, A., AND WANG, K. 2007. Mining generalized associations of semantic relations from textual web content. *IEEE Trans. Knowl. Data Eng. 19,* 2, 164–179.

JIANG, X. AND TAN, A. 2010. CRCTOL: A semantic-based domain ontology learning system. *J. Am. Soc. Inf. Sci. Technol. 61,* 1, 150–168.

KAYED, A., NIZAR, M., AND ALFAYOUMI, M. 2010. Ontology concepts for requirements engineering process in e-government applications. In *Proceedings of the 5th International Conference on Internet and Web Applications and Services (ICIW)*.

KLEIN, D. AND MANNING, C. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*.

Klien, E., Lutz, M., and Kuhn, W. 2006. Ontology-based discovery of geographic information services: An application in disaster management. *Comput. Environ. Urban Syst., 30*, 1.

Kotis, K. and Papasalouros, A. 2011. Automated learning of social ontologies. In *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances*, W. Wong, W. Liu, and M. Bennamoun, Eds. IGI Global, Hershey, PA.

Landauer, T., Foltz, P., and Laham, D. 1998. An introduction to latent semantic analysis. *Discourse Process. 25,* 1, 259–284.

Lavrac, N. and Dzeroski, S. 1994. *Inductive Logic Programming: Techniques and Applications.* Ellis Horwood, New York, NY.

Lesk, M. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th International Conference on Systems Documentation*.

Lin, D. 1994. Principar: An efficient, broad-coverage, principle-based parser. In *Proceedings of the 15th International Conference on Computational Linguistics*.

Lin, D. 1998. Dependency-based evaluation of minipar. In *Proceedings of the 1st International Conference on Language Resources and Evaluation*.

Lindberg, D., Humphreys, B., and McCray, A. 1993. The unified medical language system. *Methods Info. Med. 32,* 4, 281–291.

Linden, K. and Piitulainen, J. 2004. Discovering synonyms and other related words. In *Proceedings of the 3rd International Workshop on Computational Terminology (CompuTerm)*.

Liu, Q., Xu, K., Zhang, L., Wang, H., Yu, Y., and Pan, Y. 2008. Catriple: Extracting triples from Wikipedia categories. In *Proceedings of the 3rd Asian Semantic Web Conference (ASWC)*.

Liu, W., Jin, F., and Zhang, X. 2008. Ontology-based user modeling for e-commerce system. In *Proceedings of the 3rd International Conference on Pervasive Computing and Applications (ICPCA)*.

Liu, W., Weichselbraun, A., Scharl, A., and Chang, E. 2005. Semi-automatic ontology extension using spreading activation. *J. Univ. Knowl. Manage. 0,* 1, 50–58.

Lozano-Tello, A., Gomez-Perez, A., and Sosa, E. 2003. Selection of ontologies for the Semantic Web. In *Proceedings of the International Conference on Web Engineering (ICWE)*.

Lu, B., Tsou, B., Jiang, T., Zhu, J., and Kwong, O. 2011. Mining parallel knowledge from comparable patents. In *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances*, W. Wong, W. Liu, and M. Bennamoun, Eds. IGI Global, Hershey, PA.

Maedche, A., Pekar, V., and Staab, S. 2002. Ontology learning part one—on discovering taxonomic relations from the Web. In *Web Intelligence*, N. Zhong, J. Liu, and Y. Yao, Eds. Springer-Verlag, Germany.

Maedche, A. and Staab, S. 2000a. Discovering conceptual relations from text. In *Proceedings of the 14th European Conference on Artificial Intelligence*.

Maedche, A. and Staab, S. 2000b. The Text-to-Onto ontology learning environment. In *Proceedings of the 8th International Conference on Conceptual Structures*.

Maedche, A. and Staab, S. 2001. Ontology learning for the Semantic Web. *IEEE Intell. Syst. 16,* 2, 72 –79.

Maedche, A. and Staab, S. 2002. Measuring similarity between ontologies. In *Proceedings of the European Conference on Knowledge Acquisition and Management (EKAW)*.

Maedche, A. and Volz, R. 2001. The ontology extraction & maintenance framework: Text-to-Onto. In *Proceedings of the IEEE International Conference on Data Mining*.

Massey, L. and Wong, W. 2011. A cognitive-based approach to identify topics in text using the Web as a knowledge source. In *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances*, W. Wong, W. Liu, and M. Bennamoun, Eds. IGI Global, Hershey, PA.

Medelyan, O. and Witten, I. 2005. Thesaurus-based index term extraction for agricultural documents. In *Proceedings of the 6th Agricultural Ontology Service (AOS)*.

Mihalcea, R. and Csomai, A. 2005. Senselearner: Word sense disambiguation for all words in unrestricted text. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics*.

Mika, P. 2007. Ontologies are us: A unified model of social networks and semantics. *J. Web Semant. 5,* 1, 5–15.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. 1990. Introduction to WordNet: An on-line lexical database. *Int. J. Lexicography 3,* 4, 235–244.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP)*.

Missikoff, M., Navigli, R., and Velardi, P. 2002. Integrated approach to Web ontology learning and engineering. *IEEE Comput. 35,* 11, 60–63.

Navigli, R. and Velardi, P. 2002. Semantic interpretation of terminological strings. In *Proceedings of the 3rd Terminology and Knowledge Engineering Conference*.

Neumann, G., Backofen, R., Baur, J., Becker, M., and Braun, C. 1997. An information extraction core system for real world german text processing. In *Proceedings of the 5th International Conference of Applied Natural Language*.

O'Hara, T., Mahesh, K., and Nirenburg, S. 1998. Lexical acquisition with WordNet and the microkosmos ontology. In *Proceedings of the Coling-ACL Workshop on Usage of WordNet in Natural Language Processing Systems*.

Oliveira, A., Pereira, F., and Cardoso, A. 2001. Automatic reading and learning from text. In *Proceedings of the International Symposium on Artificial Intelligence (ISAI)*.

Park, H., Kwon, S., and Kwon, H. 2009. Ontology-based approach to intelligent ubiquitous tourist information system. In *Proceedings of the 4th International Conference on Ubiquitous Information Technologies & Applications (ICUT)*.

Pedersen, T., Patwardhan, S., and Michelizzi, J. 2004. WordNet:similarity: Measuring the relatedness of concepts. In *Proceedings of the Demonstration Papers at the Conference of the North American Chapter of the Association for Computational and Linguistics: Human Language Technologies (HLT-NAACL)*.

Pei, M., Nakayama, K., Hara, T., and Nishio, S. 2008. Constructing a global ontology by concept mapping using Wikipedia thesaurus. In *Proceedings of the 22nd International Conference on Advanced Information Networking and Applications*.

Pereira, F. and Cardoso, A. 1999. Dr. Divago: Searching for new ideas in a multi-domain environment. In *Proceedings of the 8th Cognitive Science of Natural Language Processing (CSNLP)*.

Pereira, F., Oliveira, A., and Cardoso, A. 2000. Extracting concept maps with Clouds. In *Proceedings of the Argentine Symposium of Artificial Intelligence (ASAI)*.

Ponte, J. and Croft, B. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st International Conference on Research and Development in Information Retrieval*.

Porzel, R. and Malaka, R. 2004. A task-based approach for ontology evaluation. In *Proceedings of the 16th European Conference on Artificial Intelligence*.

Raskin, R. and Pan, M. 2005. Knowledge representation in the Semantic Web for earth and environmental terminology (sweet). *Comput. Geosci. 31,* 9, 1119–1125.

Resnik, P. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Arti. Intell. Res. 11,* 1, 95–130.

Ribas, F. 1995. On learning more appropriate selectional restrictions. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*.

Sabou, M., Wroe, C., Goble, C., and Mishne, G. 2005. Learning domain ontologies for Web service descriptions: An experiment in bioinformatics. In *Proceedings of the 14th International Conference on the World Wide Web*.

Salton, G. and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Info. Process. Manag. 24,* 5, 513–523.

Sanchez, D. and Moreno, A. 2008. Learning non-taxonomic relationships from Web documents for domain ontology construction. *Data Knowl. Eng. 64,* 3, 600–623.

Schmid, H. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Sclano, F. and Velardi, P. 2007. Termextractor: A Web application to learn the shared terminology of emergent Web communities. In *Proceedings of the 3rd International Conference on Interoperability for Enterprise Software and Applications (I-ESA)*.

Senellart, P. and Blondel, V. 2003. Automatic discovery of similar words. In *Survey of Text Mining*, M. Berry, Ed. Springer-Verlag, Berlin.

Shamsfard, M. and Barforoush, A. 2003. The state of the art in ontology learning: A framework for comparison. *Knowl. Eng. Rev. 18,* 4, 293–316.

Shamsfard, M. and Barforoush, A. 2004. Learning ontologies from natural language texts. *Int. J. Human Comput. Stud. 60,* 1, 17–63.

Sleator, D. and Temperley, D. 1993. Parsing english with a link grammar. In *Proceedings of the 3rd International Workshop on Parsing Technologies*.

Snow, R., Jurafsky, D., and Ng, A. 2005. Learning syntactic patterns for automatic hypernym discovery. In *Proceedings of the 17th Conference on Advances in Neural Information Processing Systems*.

SNOW, R., JURAFSKY, D., AND NG, A. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the ACL 23rd International Conference on Computational Linguistics (COLING)*.

SOMBATSRISOMBOON, R., MATSUO, Y., AND ISHIZUKA, M. 2003. Acquisition of hypernyms and hyponyms from the WWW. In *Proceedings of the 2nd International Workshop on Active Mining*.

SPECIA, L. AND MOTTA, E. 2006. A hybrid approach for relation extraction aimed at the Semantic Web. In *Proceedings of the International Conference on Flexible Query Answering Systems (FQAS)*.

SPILIOPOULOU, M., RINALDI, F., BLACK, W., AND PIERO-ZARRI, G. 2004. Coupling information extraction and data mining for ontology learning in parmenides. In *Proceedings of the 7th International Conference on Computer-Assisted Information Retrieval (RIAO)*.

SRIKANT, R. AND AGRAWAL, R. 1997. Mining generalized association rules. *Future Gen. Comput. Syst. 13,* 2-3, 161–180.

STREHL, A. 2002. Relationship-based clustering and cluster ensembles for high-dimensional data mining. Ph.D. dissertation, University of Texas at Austin.

TANG, J., LEUNG, H., LUO, Q., CHEN, D., AND GONG, J. 2009. Towards ontology learning from folksonomies. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI)*.

TURCATO, D., POPOWICH, F., TOOLE, J., FASS, D., NICHOLSON, D., AND TISHER, G. 2000. Adapting a synonym database to specific domains. In *Proceedings of the ACL Workshop on Recent Advances in Natural Language Processing and Information Retrieval*.

TURNEY, P. 2001. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning (ECML)*.

USCHOLD, M. AND GRUNINGER, M. 2004. Ontologies and semantics for seamless connectivity. *ACM SIGMOD 33,* 4, 58–64.

VARGAS-VERA, M., DOMINGUE, J., KALFOGLOU, Y., MOTTA, E., AND SHUM, S. 2001. Template-driven information extraction for populating ontologies. In *Proceedings of the IJCAI Workshop on Ontology Learning*.

VELARDI, P., FABRIANI, P., AND MISSIKOFF, M. 2001. Using text processing techniques to automatically enrich a domain ontology. In *Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS)*.

VELARDI, P., NAVIGLI, R., CUCCHIARELLI, A., AND NERI, F. 2005. Evaluation of OntoLearn, a methodology for automatic learning of ontologies. In *Ontology Learning from Text: Methods, Evaluation and Applications*, P. Buitelaar, P. Cimmiano, and B. Magnini, Eds. IOS Press, Hershay, PA.

VITANYI, P., BALBACH, F., CILIBRASI, R., AND LI, M. 2009. Normalized information distance. In *Information Theory and Statistical Learning*, F. Emmert-Streib and M. Dehmer, Eds. Springer, New York, NY.

VOLKER, J., FERNANDEZ-LANGA, S., AND SURE, Y. 2008. Supporting the construction of Spanish legal ontologies with Text2Onto. In *Computable Models of the Law*, P. Casanovas, G. Sartor, N. Casellas, and R. Rubino, Eds. Springer-Verlag, Berlin, Heidelberg.

VRONIS, J. AND IDE, N. 1998. Word sense disambiguation: The state of the art. *Comput/ Ling. 24,* 1, 1–41.

WEBER, N. AND BUITELAAR, P. 2006. Web-based ontology learning with ISOLDE. In *Proceedings of the ISWC Workshop on Web Content Mining with Human Language Technologies*.

WEICHSELBRAUN, A., WOHLGENANNT, G., AND SCHARL, A. 2010. Augmenting lightweight domain ontologies with social evidence sources. In *Proceedings of the 9th International Workshop on Web Semantics*.

WERMTER, J. AND HAHN, U. 2005. Finding new terminology in very large corpora. In *Proceedings of the 3rd International Conference on Knowledge Capture*.

WONG, W. 2009. Learning lightweight ontologies from text across different domains using the Web as background knowledge. Ph.D. dissertation, University of Western Australia.

WONG, W., LIU, W., AND BENNAMOUN, M. 2011. Constructing specialised corpora through analysing domain representativeness of websites. *Lang. Resources Eval. 45*, 2, 209–241.

WONG, W., LIU, W., AND BENNAMOUN, M. 2007. Tree-traversing ant algorithm for term clustering based on featureless similarities. *Data Mining Knowl. Discov. 15,* 3, 349–381.

WONG, W., LIU, W., AND BENNAMOUN, M. 2008a. Constructing Web corpora through topical Web partitioning for term recognition. In *Proceedings of the 21st Australasian Joint Conference on Artificial Intelligence (AI)*.

WONG, W., LIU, W., AND BENNAMOUN, M. 2008b. Determination of unithood and termhood for term recognition. In *Handbook of Research on Text and Web Mining Technologies*, M. Song and Y. Wu, Eds. IGI Global, Hershey, PA.

WONG, W., LIU, W., AND BENNAMOUN, M. 2009a. Acquiring semantic relations using the Web for constructing lightweight ontologies. In *Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*.

WONG, W., LIU, W., AND BENNAMOUN, M. 2009b. A probabilistic framework for automatic term recognition. *Intell. Data Anal. 13,* 4, 499–539.

YANG, Y. AND CALMET, J. 2005. OntoBayes: An ontology-driven uncertainty model. In *Proceedings of the International Conference on Intelligent Agents, Web Technologies and Internet Commerce (IAWTIC)*.

YANGARBER, R., GRISHMAN, R., AND TAPANAINEN, P. 2000. Automatic acquisition of domain knowledge for information extraction. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*.

YEH, J. AND YANG, N. 2008. Ontology construction based on latent topic extraction in a digital library. In *Proceedings of the 11th International Conference on Asian Digital Libraries (ICADL)*.

ZELLE, J. AND MOONEY, R. 1993. Learning semantic grammars with constructive inductive logic programming. In *Proceedings of the 11th National Conference of the American Association for Artificial Intelligence (AAAI)*.

ZHANG, Z. AND CIRAVEGNA, F. 2011. Named entity recognition for ontology population using background knowledge from Wikipedia. In *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances*, W. Wong, W. Liu, and M. Bennamoun, Eds. IGI Global, Hershey, PA.

ZHOU, L. 2007. Ontology learning: State of the art and open issues. *Info. Technol. Manage. 8,* 3, 241–252.

Received October 2010; revised February 2011; accepted March 2011