Journal of Biomedical Informatics 44 (2011) 163–179

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Methodological Review

Natural Language Processing methods and systems for biomedical ontology learning

Kaihong Liu [a], William R. Hogan [b], Rebecca S. Crowley [a,c,*]

[a] Department of Biomedical Informatics, University of Pittsburgh School of Medicine, USA, USA
[b] Division of Biomedical Informatics, University of Arkansas for Medical Sciences, USA
[c] Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA

ARTICLE INFO

Article history:
Received 21 April 2009
Available online 18 July 2010

Keywords:
Ontology
Ontology learning from text
Ontology enrichment
Information extraction
Natural Language Processing

ABSTRACT

While the biomedical informatics community widely acknowledges the utility of domain ontologies, there remain many barriers to their effective use. One important requirement of domain ontologies is that they must achieve a high degree of coverage of the domain concepts and concept relationships. However, the development of these ontologies is typically a manual, time-consuming, and often error-prone process. Limited resources result in missing concepts and relationships as well as difficulty in updating the ontology as knowledge changes. Methodologies developed in the fields of Natural Language Processing, information extraction, information retrieval and machine learning provide techniques for automating the enrichment of an ontology from free-text documents. In this article, we review existing methodologies and developed systems, and discuss how existing methods can benefit the development of biomedical ontologies.

Published by Elsevier Inc.

1. Background

1.1. Knowledge resources used in Natural Language Processing

Natural Language Processing (NLP) and text mining are research fields aimed at exploiting rich knowledge resources with the goal of understanding, extraction and retrieval from unstructured text. Knowledge resources that have been used for these purposes include the entire range of terminologies, including lexicons, controlled vocabularies, thesauri, and ontologies. For the purposes of this description we follow the framework for describing terminologies and terminological systems defined by de Keizer [1,2] and Cornet [3]. The authors define concepts as "cognitive constructs" of objects that are built using the "characteristics of the objects", terms as "language labels" for concepts, and synonyms as two or more terms that designate "a unique concept."

For simple NLP tasks, such as named entity recognition, almost any type of terminology can be used. Slightly more complex tasks such as identification of concepts, requires the representation of synonyms, and therefore limits the resources to terminological systems such as controlled vocabularies and ontologies that encode multiple lexical representations in natural language [4]. For example, "liver cell" and "hepatocyte" would be represented in the vocabulary or ontology as synonyms, and therefore during NER they would be classified as the same concept.

In contrast, some NLP tasks require more complex relationships between concepts, and therefore limit the types of terminological systems that may be used. Examples include word sense disambiguation [5], co-reference resolution [6–8], and discourse reasoning and extraction of attributes and values [9]. For example, if "hepatocellular carcinoma" and "liver neoplasm" are both used in a document to refer to the same entity, then these terms can be determined to co-refer if a relationship is represented in the terminology [10].

Ontologies can be used to make even more complex inferences and to derive rules necessary for semantic interpretation [11,12] and question and answering systems [13]. For this reason, ontologies have been of particular interest to researchers developing NLP systems. For example, to answer the question: "What role do infectious organisms play in liver cancer?" an ontology can be used to perform the query expansion and retrieve related textual information, if it contains the following information: (1) a synonym relationship between 'liver cancer' and 'hepatocellular carcinoma', (2) a hierarchical relationship between various hepatitis viruses and 'infectious organism', (3) an etiologic relationship between hepatitis viruses and hepatocellular carcinoma.

1.2. Ontologies and ontology development

Researchers define 'ontology' in different ways [14–17], but these definitions have in common that an ontology is a representation of entities and their relationships in a particular domain,

* Corresponding author. Address: Department of Biomedical Informatics, UPMC Cancer Pavilion, Suite 301, 5150 Centre Avenue, Pittsburgh, PA 15232, USA. Fax: +1 412 647 5380.
E-mail address: mailto: kaihong@pitt.edu (R.S. Crowley).

1532-0464/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.jbi.2010.07.006

debates to whether the 'entities' represented are concepts [18] or real-world things [19] notwithstanding. A key requirement is that each entity has one unique reference in the ontology, typically a meaningless identifier to avoid confusion among natural language terms. Each identifier is linked to at least one natural language term, and often to greater than one natural language term to capture the synonymy inherent in human language. A standard ontology facilitates aggregation of data from multiple sources if each source uses the identifiers from the ontology. Interoperability is one of the primary—if not the primary—reason that groups have been engaged in the development of ontologies.

Ontology developers usually capture the relationships among entities as formal, logical relationships. To do so, they frequently use one out of a family of logics known as description logics. Description logics constitute a family of fragments of first-order logic (nearly all of which are decidable), in which members of this family are primarily differentiated based on the set of allowed logical operators (for example, some exclude negation and universal quantification), which in turn determines the computational complexity of inference with the language. The Web Ontology Language (OWL) is a standard ontology language that captures the semantics of many description-logic languages.

A key consideration for NLP is that an ontology be complete with respect to the entities represented as well as their relationships and natural-language synonyms. To return to our example, to retrieve documents that discuss hepatocellular carcinoma, an NLP system requires an ontology that has an identifier for hepatocellular carcinoma, links from that identifier to the natural language terms 'hepatocellular carcinoma', 'liver cancer', "HCC", etc., and relationships of that identifier to identifiers for other entities, such as *hepatocellular carcinoma is a liver neoplasm*. It follows that when an ontology lacks a representation of an entity, a particular term for it, or some of its particular relationships, the quality of NLP based solely on that ontology will suffer. Lack of any representation of an entity inhibits detection of that entity. Lack of a synonym prevents recognition of the entity when a document uses the synonym to refer to it. Lack of a relationship might prevent finding answers to such questions as *What role do infectious organisms play in liver cancer?*

At present, the process of ontology development is largely manual. Humans must add identifiers and their synonyms and relationships one by one. The investment in ontology development is huge. The National Human Genome Research Institute has funded the gene ontology (GO) Consortium since 2001 [20], when the GO was already enjoying widespread success. In 2009, this funding was $3.4 million plus a $1 million supplement [20]. In 2005, the National Center for Biomedical Ontology (NCBO) received $18.8 million over 5 years [21]. An effort to build the infectious disease ontology just received $1.25 million over 4 years [22]. The National Science Foundation recently invested >$900,000 over 2 years in an ontology of *Hymenoptera* [23]. The National Library of Medicine has paid approximately $6 million per year for the ongoing development of SNOMED-CT since 2007 [24], after an initial investment of $32.4 million in 2003 [25].

One approach to facilitating this manual process is to use informatics tools to accelerate the interactions among domain experts and ontologists necessary to the ontology development process. An important recent development is the NCBO's BioPotal. BioPortal enables the biomedical community to find, comment on, and contribute to biomedical ontologies, thereby facilitating interactions among ontology users and developers to increase the value of the ontologies [26]. Stanford has developed Collaborative Protégé to allow collaborative ontology development in real time by users in different locations [27]. The earliest examples of such technologies date to the mid-1990s with work done by Campbell et al. to facilitate geographically-distributed development of SNOMED-RT and its successor SNOMED-CT [28].

Another approach to reducing resources required is division of labor. Put simply, the goal is to avoid the wastefulness of recreating multiple representations of the same entity (and its synonyms and relationships) in multiple ontologies, which results in multiple identifiers for entities, one per ontology. The OBO Foundry seeks to avoid this problem and thereby facilitate ontology development by mandating orthogonality of ontologies. That is, it has a well-defined goal of having only one representation of an entity in all of the ontologies in the Foundry [29]. Already, per Smith et al., this principle has resulted in the consolidation of several ontologies [29]. This approach also has the goal of increasing interoperability by avoiding the necessity for 'mapping' identifiers among ontologies that represent the same entities (i.e., asserting that identifiers from multiple ontologies refer to the same entity).

Lastly, there is a large body of research on automating the development and maintenance of ontologies using NLP. Because literature and text documents are major mechanisms for reporting new knowledge about a domain, ontological knowledge is often stated explicitly or implicitly within the text, and these reference documents serve as important knowledge-rich resources for ontology learning. As NLP often uses ontological knowledge to interpret the texts (see Section 1.1), NLP can also help to enrich and enhance the linguistic realization of ontology. Therefore, many researchers have been utilizing methods from fields of NLP, computational linguistics (CL) and artificial intelligence (AI) to partially or fully automate semantic knowledge extraction. This approach has been termed "ontology learning", and represents a sub-field of knowledge acquisition (KA). The goal of this paper is to survey these methods.

1.3. Ontology learning and ontology learning tasks

Knowledge acquisition (KA) is a broad field that encompasses the processes of extracting, creating, structuring knowledge from heterogeneous resources, including experts [30]. Semi-automated and automated approaches to KA utilize data that may be derived from structured, semi-structured, or unstructured data sources, and may result in knowledge representations of many different types [31]. Ontology learning (OL), however, is limited to the extraction of ontological elements from knowledge-rich resources. A further delineation is made for ontology learning from text, which builds on a large body of work within the fields of NLP, CL and AI [32,33]. In biomedicine, text resources for ontology learning from text include the scientific literature and clinical documents, many of which are already available in electronic format. Finally, ontology learning from text can be further subdivided by task based on the ontological element that is learned from the resources [32]. These tasks include term extraction, synonym extraction, concept extraction (both taxonomic and non-taxonomic), relationship extraction and axiom extraction (an axiom is defined as a rule that is used to constrain the information in an ontology).

The purpose of this paper is to review research on ontology learning from text, both within and outside of biomedical informatics. Because the potential breadth of this review is very large, we have made the following decisions and definitions in limiting our scope:

(1) Although there continues to be dissent over whether instances (individuals) should be included in biomedical ontologies at all, many NLP tasks including information extraction, co-reference resolution and question answering cannot be accomplished without knowledge of instances and their relationship to the corresponding ontology classes. Many researchers in KA and OL consider learning of new ontology instances to be part of ontology learning [32], as it can be encompassed by some combination of term extraction, synonym extraction and concept extraction, depending

on how knowledge is modeled in the ontology. For these reasons, we choose to define instance learning as a task of ontology learning in this review. We recognize that this task may not be relevant to all ontology engineering efforts.

(2) As we have previously described, the broader field of KA includes research that is easily applied to some of these tasks (particularly term and synonym extraction). Therefore, for these tasks we have not strictly limited our review to those methods that could be labeled as "ontology learning". For a more complete treatment of the general field of KA, and automated approaches, the reader is referred to recent review articles and book chapters [30,34–36].

(3) We have chosen to exclude axiom learning from the ontology learning tasks reviewed, because there has been so little relevant work in this area.

1.4. NLP approaches to ontology learning

For the past several decades, fields of studies such as computational linguistics, NLP, machine learning (ML), and AI have developed methods and algorithms for information retrieval and extraction from free-text knowledge resources. Some of these methods have been used and tested for ontology learning from text and have shown promising results. In general, these methods can be categorized into symbolic, statistical, and hybrid approaches (Table 1). The symbolic approach utilizes linguistic information to extract information from text. For example, noun phrases are considered to be lexicalized concepts and are often used to represent concepts in an ontology. Linguistic rules describing the relationships between terms in the text can also be used to identify conceptual relationships within an ontology. The most common symbolic approach is to use lexico-syntactic pattern (LSP) matching, which was first explored by Hearst [37]. LSPs are surface relational markers that exist in a natural language. For example, in the phrase "systemic granulomatous diseases such as Crohn's disease or sarcoidosis" the words "such as" can help us infer that "systemic granulomatous diseases" is a hypernym of "Crohn's disease" and "sarcoidosis". Another symbolic approach is to use the internal syntactic structure of component terms. Concepts are often represented using compound, multi-word terms. In general, a compound term is more specific than a single compositional term. The basis of this method is the assumption that a compound term is likely a hyponym of a single term. For example, using this approach the term "prostatic carcinoma" can be considered to be a hyponym of "carcinoma". It is also possible to use multiple symbolic approaches at the same time, for example the LSP method can be used with information from compound terms.

Table 1
Ontology learning tasks and their corresponding learning methods.

Task	Primary Method	Secondary Method	Authors
Synonym and concept extraction			
	Symbolic	Compound noun information	Hamon [62]
		Lexico-syntactic patterns (LSP)	Downey [63]
		LSP + compound noun information	Moldovan, Girju [64]
	Statistical		Church [42]
			Smadia [43]
		Clustering	Grefenstette [65], Hindel [46]
			Geffet, Dagan [66]
			Agirre [48]
			Faatz, Steimetz [67]
			Collier [52]
		Hidden Markov Model (HMM)	Bikel [68]
			Morgan [53]
			Shen [54]
		Support Vector Model (SVM)	Kazama [55]
			Yamamoto [56]
		Conditional Random Fields (CRFs)	Chanlekha [69]
Taxonomic relationship extraction			Hearst [37]
			Caraballo [70]
	Symbolic	LSP	Cederberg, Widdow [71]
			Fiszman [72]
			Snow [73]
			Riloff [74]
		Compound noun information	Velardi [75] Cimiano [76]
			Rinaldi [77]
			Morin [78]
			Bodenreider [79]
			Ryu [80]
	Statistical	Clustering	Alfonseca, Manandler [58]
		Machine learning	Witschel [81]
Non-taxonomic relationship extraction			Berland [82]
	Symbolic	LSP	Sundblad [83]
			Girju [84]
			Nenadić, Ananiadou [85]
	Statistical	Co-occurring information	Kavalec [86]
			Gulla [49]
		Association rule mining	Chefi [50]
			Bodenreider [51]
Ontology generation (combining all tasks)	Statistical	Dependency triples	Lin [45]
		Nearest neighbor clustering	Blaschke, Valencia [87]

The statistical approach uses large corpora of text data, so this approach has also been characterized as the "corpus-based approach". Harris [38] popularized this approach with his distributional hypothesis, advancing Firth's notion that "a word is characterized by the company it keeps" [39]. Building on Harris's theory, it became common practice to classify words not only on the basis of their meaning, but also on the basis of their co-occurrence with other words. The advantage of this method is that it requires minimal prior knowledge and can be generalized to other domains. However, for reliable information to be obtained, a large corpus of text is needed. Statistical techniques often utilize different linguistic principles and features for statistical measurements to extract semantic information. One of these linguistic principles is selectional restrictions [40], in which syntactic structures provide relevant information about semantic content.

Statistical methods can be categorized into two major categories: clustering approaches and machine learning methods. The clustering technique for extraction is based on a similarity measure, whereas the machine learning technique attempts to treat the knowledge extraction problem as a classification process. Clustering is useful for two purposes. First, the similarity measurements can provide information about the hierarchical relationships of concepts (relationship extraction). Second, the identification of distinct clusters of similar terms can help in identifying concepts and their synonyms.

The extraction techniques for clustering similar terms are based on definitions of a context within a given corpus. In general, the context of the target word refers to the surrounding linguistic elements. However, the precise definition of context can vary somewhat depending on the scope. For example, the "first order word context" defined by Grefenstette [41], utilizes information only in the immediate vicinity of the target word [42,43]. In contrast, the "second order word context" utilizes syntactic information, such as noun-modifiers [44], dependency triples [45], verb-arguments [46], and preposition structures [41,47]. When utilizing second order context similarity to cluster similar words, we would expect semantically similar words to cluster even though they would not typically appear next to each other. For example, the synonyms 'tumor' and 'tumour' would cluster together because they are likely to appear in similar contexts, even though they would not be found together. The context can be further defined as the entire document. In this approach, concepts are represented by a vector of co-occurring terms within a set of domain-specific documents, as a concept signature [45,48]. Similarity between concepts can then be calculated by comparing concept signatures. Another approach that utilizes the context of the entire document is the association rule mining technique for concept relationship discovery [49–51].

Although machine learning is the major approach used for many NLP tasks such as POS tagging, chunking and co-reference resolution, most applications of machine learning to ontology learning from text focus on the relatively simpler task of new concept identification and use supervised methods [52–56]. Using machine learning methods to identify the precise taxonomic location for a concept is a much more difficult task for fully automated systems [57–60].

Despite the widely accepted belief that statistical methods for ontology learning provide better coverage and scalability than symbolic methods, Basili [61] points out that statistical methods only provide a probability. The output is often represented by words, word strings or word clusters with associated probabilities. The conceptual explanation of the results is not provided. Ultimately, a human analyst must make sense of this data, because, at present, full automation seems unachievable. Therefore, many researchers have explored the potential of combining the statistical and the symbolic approaches for knowledge extraction.

The remainder of this paper is organized as follows. First, we review the methods and algorithms that have been used for ontology learning (Section 3) categorized by ontology learning task and by approach (Table 1). For each of these categories, we review related papers that are prominent in the field of ontology learning, focusing on algorithmic methods, and describe the advantages and disadvantages of each method. Second, we provide examples of several state-of-the art systems that use these various approaches to support ontology learning from text (Section 4). Third, we discuss how these techniques could be used to develop more sophisticated methods and systems for biomedical ontology learning, as well as the barriers that may impede such progress (Section 5).

2. Retrieval and selection of articles considered in this review

Articles were retrieved using three approaches: (1) search of references across multiple WWW sources using the Google search engine (2) review of a recently published book containing chapters relevant to this subject, and (3) iterative review of related citations.

For the internet search, we used the key words "ontology learning from text", "ontology enrichment", and "NLP and Ontology development" to retrieve research articles from multiple sources. From all articles returned, we included articles relevant to the topic, with either high search engine ranking (presence within the first 100 items) or greater than 15 citations on CiteSeer. We also included articles cited in the book "Ontology learning from text: methods, evaluation and application" [32]. Finally, we iteratively reviewed citations from these sources to find other relevant articles, and then reviewed the references from the newly identified articles. Using this process, we read and considered a total of 343 articles, of which 150 are discussed in this review paper. Of these 150 articles, 51 are discussed in detail as exemplars of the various approaches.

3. Research on ontology learning from text

We review approaches to ontology learning from text, based on their associated learning task: synonym and concept extraction (Section 3.1), taxonomic relationship extraction (Section 3.2), non-taxonomic relationship extraction (Section 3.3), and generation of ontologies de novo (Section 3.4). We consider the task of term extraction (instance extraction) to be encompassed by concept or synonym extraction, and it is therefore not separately considered. In many cases, a particular method can be used for more than one task, which is particularly common among the statistical methods. For the purposes of this review, we have classified each paper by the task that we consider most salient, and noted other tasks that may be accomplished when relevant. Because we focus on describing approaches and algorithms, we have further distinguished approaches that are primarily symbolic from those which are primarily statistical, and by primary methodology type (e.g. LSP, clustering), noting those cases in which the approaches overlap.

3.1. Extraction of synonyms and concepts

Extraction of synonyms and concepts has been approached using a variety of methods. In many cases, a particular method cannot distinguish between these ontological elements. In other cases, a particular method that has been used for one of these tasks could easily be used for another learning task. Thus, we consider approaches in this category along a spectrum of complexity, starting with symbolic methods designed primarily to extract synonyms.

3.1.1. Symbolic methods

Compound noun information provides a simple symbolic method for synonym identification. Hamon et al. [62] used a general purpose thesaurus as the knowledge resource along with the following three heuristics: (1) *IF two compound terms' noun heads which are identical and have modifiers which are synonyms; or (2) IF two noun heads are synonyms and have modifiers which are identical; or (3) IF two noun heads are synonyms and have modifiers which are also synonyms, THEN the two compound terms are synonyms.* Using a biomedical example, the terms "hepatic tumor" and "hepatic tumour" can be considered synonyms because the modifiers are identical and the head nouns "tumor" and "tumour" are synonyms. Working with a corpus of engineering documents, Hamon et al. evaluated this method and found that 37% of the extracted synonym pairs were correct. The first two heuristics were most effective, producing 95% of the total correct synonyms.

Another approach for extracting synonyms and concepts relies on the use of lexico-syntactic patterns (LSP), often using a bootstrap method. In this case, a set of seed concepts or patterns is used to extract new concepts or patterns, initiating a cycle of discovery and extraction. An important problem is to control the quality of the extraction, using some discriminating performance metric. Downey and colleagues [63] exemplify this approach, which they defined as the pattern learning algorithm (PL). The algorithm started with a set of seed instances generated by domain-independent patterns (e.g. Hearst patterns). For each seed word in the set, they retrieved more instances that contained the seed word from the WWW. Patterns were obtained by creating a window of w words around the seed word (w was set to 4 in their experiment), which acted as a threshold for selecting pattern candidates. In the first step, patterns with relatively high estimated recall and precision were selected, and these patterns were used to extract new concept candidates from the WWW in order to improve the recall. Using the selected patterns boosted recall from 50% to 80%. In the second step, they used Turney's [88] point wise mutual information (PMI), in order to improve the precision. PMI is a statistical measure of the strength of association between an extraction and discriminator (pattern). PMI is calculated as Counts (D + E)/Counts (E) where D is the pattern, E is the extraction and D + E is the pattern with the extraction as the instance placeholder. Downey and colleagues used the PMI scores for a given extraction as features in a Naïve Bayes classifier, to determine whether the pattern should be used as an extractor. For example, in the pattern "city of <CITY>" D represents the pattern "City of <X>", while E represents the various instances extracted as <CITY>. This pattern has a high PMI because "City of" rarely extracts instances that are not cities, and the cities extracted are frequently associated with this pattern. In contrast, the pattern "<CITY> hotels" has a low PMI because many other terms (such as "budget") are also extracted. The classification step is performed to improve accuracy because a single threshold will not work for every domain. Using this method of discrimination, Downey increased precision from 70% to 87%. This method seems highly amenable to applications in the biomedical domain as we often observe patterns that have high PMIs. For example "<protein> activates <X>" will extract either a "Protein" or "Process" in biomedical domain (e.g. "Fyn activates Cbl", "Bcl-2 activates apoptosis"). The method could be used to extract terms which may be either new synonyms or new concepts, but it is unlikely to distinguish between them.

Combining both compound noun information and lexico-syntactic pattern matching (LSP), Moldovan and Girju [64] developed an approach to enrich domain-specific concepts and relationships in WordNet. The source for acquiring new knowledge was a general English corpus and was augmented by using other lexical resources such as domain-specific corpora and general dictionaries. The user provided domain-specific seed concepts, which were used to discover new concepts and relations from the source. The method was tested on five seed concepts selected from the financial domain: "interest rate", "stock market", "inflation", "economic growth", and "employment". Queries were formed with each of these concepts and a corpus of 5000 sentences was extracted automatically from the Internet and TREC-8 corpora. From these, they discovered a total of 264 new concepts not defined in WordNet, of which 221 contain the seeds and 43 are other related concepts. Compound noun information and LSP can also be used to extract taxonomic relationships. In the case of Moldovan and Girju, they used this combined method to discover 64 new relationships that link these concepts with each other or with other either WordNet concepts.

3.1.2. Statistical methods

3.1.2.1. Methods that use clustering approaches. Clustering methods have been commonly applied to concept and synonym extraction, because text corpora provide a great deal of data for computing similarity measures. These methods may be able to distinguish synonyms from new concepts based on the degree of statistical similarity. Because these measures can be compared to the existing ontology, these methods can also be used to suggest placement of the concept in the hierarchy.

One of the first to suggest the clustering approach was Church [42] who proposed methods to measure word association based on the information theoretic notion of mutual information. The association ratio of two words (x, y) was calculated as the probability of observing x and y together (the joint probability) divided by the probability of observing x and y independently (the product of the marginal probabilities). If there is a genuine association between x and y, then the joint probability $P(x, y)$ should be larger than chance $P(x)\,P(y)$. In this case, context is the immediate vicinity of a given word in a window. Church suggested that smaller window sizes might identify fixed expressions (idioms) and other relationships that hold over short ranges, while larger window sizes might highlight semantic concepts and other relationships over a larger scale.

Smadia [43] further extended Church's proposal by using Church's method as the first stage and adding two more stages to raise the precision. The two added stages are both filtering functions. One of them calculated the histogram of the frequency of the target word (x) relative to position of the collocated word (y) with a five word window before and after the target. If the histogram was flat, the association between x and y was rejected. The other filter calculated which spike to pick if more than one spike appeared in the histogram. These two additional functions eliminated the noise introduced by non-specific associations.

A similar approach is described in Grefenstette [65] and Hindle [46], both of whom describe the clustering of terms according to the verb-argument structures they display in the text corpus. The approach termed "selectional restriction" exploits the restrictions on what words can appear in a specific structure. For example, wine might be "drunk", "produced", or "sold", but not "pruned". Using 6 million words in the 1987 AP news corpus, Hindle extracted a set of Subject–Verb–Object triples and calculated the mutual information between verb-noun pairs. Using this approach, nouns with the highest associations as objects of the verb "drink" were "beer", "tea", "Pepsi", "wine", "water", etc. Then, they calculated the similarity between nouns by considering how much mutual information these nouns shared with the verbs in the corpus. This phenomenon may be even more pronounced in biomedical domains, in keeping with Harris's sublanguage theory [89,90] as meanings of a term and vocabularies are further restricted. For example, in the biomolecular domain, the predicate "INTERACTION" includes two subcategories, "activate" and "attach". For semantic groups "protein" and "process", "protein" is constrained

to co-occur with the "activates process", not the "attaches process" pattern. Therefore, the Subject–Verb–Object triple approach may prove to be very effective for similar-term extraction. Examples of the effective use of this technique in biomedical domains include Friedman's MedLEE [91] and Sager's Linguistic String Project (LSP) system [92].

Geffet and Dagan [66] further explored the relationship between the distributional characterization of words. They proposed two new hypotheses as a refinement to the distributional similarity hypothesis. They claimed that distributional similarity captures a somewhat loose notion of semantic similarity. But in the case of tight semantic relationships, for example synonym relationships, the distributional similarity measure may not be sufficient. In this work, they paid particular attention to this type of semantic relationship. They describe a "lexical entailment relationship" as a relationship between a pair of words such that the meaning of one word sense can be inferred by substitution with the paired word. The refined versions of the distributional similarity hypothesis for lexical entailment inference are as follows: Let v_i and w_j be two word senses of the words v and w, correspondingly, and let $v_i \geqslant w_j$ denote the (directional) entailment relation between the two words senses. Also they assume that they have a measure that determines the set of characteristic features for meaning of each word sense. Then (1) If $v_i \geqslant w_j$ then all the characteristics (syntactic) of v_i are expected to appear with w_j. (2) If all the characteristic features (syntactic based) of v_i appear with w_j then we expect that $v_i \geqslant w_j$. They performed an empirical analysis on a selected test sample to test the validity of the two distributional inclusion hypotheses. The first hypothesis completely fit the data while the second hypothesis held 70% of the time. They further employed the inclusion hypotheses as a method to filter out non-entailing word pairs. Precision was improved by 17% and F1 was improved by 15% over the baseline.

By incorporating information from the entire document, Agirre [48] exploited a topic signature approach for concept clustering to enrich WordNet. He showed that topic signatures could be used to disambiguate word senses, a common problem in using text corpora for ontology learning. His work followed Lin and Hovy [45], who originally developed this approach for text summarization. First, he composed a query using the WordNet concept with its synset to extract documents from the WWW. Each document collection was used to build a topic signature for each concept sense. The topic signature for a concept sense, derived from WordNet, was a set of words from a collection of selected documents which had higher frequency of the concept sense when compared with the remaining documents. For a given new concept candidate, the topic signature was obtained and compared to the signature calculated for the concept sense, using the chi-square statistic. The word sense with the highest chi-square score was the chosen sense for that concept candidate.

Faatz and Steinmetz [67] developed a sophisticated method to utilize distances inherent to an existing ontology in order to optimize enrichment. The method utilized a comparison between the statistical information of word usage in a corpus and the structure of the ontology itself, based on the Kullback–Leibler divergence measure. Although they also used collocation information for the similarity measure, their method was different from those of others because they defined a parameterization by assigning weights to each word collocation feature so they could optimize the parameters used in the calculation. One interesting advantage of this approach is that it might preferentially select candidates which approximate the level of abstraction for a given ontology.

3.1.2.2. Methods that use machine learning approaches. Machine learning methods can also be used for concept and synonym extraction. Collier et al. [52] described how to extract molecular-biology terminology from MEDLINE abstracts and texts using Hidden Markov models (HMM). They trained the HMM with bigrams based on lexical and character features in a relatively small corpus of 100 MEDLINE abstracts that had been marked-up by domain experts with eleven term classes such as "proteins" and "DNA". Word features used for their HMM were based on Bikel [68] and included 23 features, such as *Digital Number, Single Capitalized Letter, Greek Letter, Capitalized and Digits, Hyphen* etc. The testing data consisted of 3300 MEDLINE abstracts from a sub domain of molecular-biology, retrieved using the query terms *human, blood cell,* and *transcription factor*. Using the HMM classifier, they extracted named entities related to the eleven classes, and determined the accuracy of classification of the named entities, using *F*-score as their metric. The method performed adequately, with an average *F*-score of 73%.

Morgan [53] further extended Collier's approach, developing a method appropriate for learning new instances without human-annotated training data. Considering such hand-annotation to be a limitation of Collier's method, Morgan leveraged an existing FlyBase resource to provide supervision. The FlyBase model database was created by human curation of published experimental findings and relations in the *Drosophila* literature. The resource contains a list of genes, related articles from which the gene entries were drawn, and a synonym lexicon. Morgan applied a simple pattern matching method to identify gene names in the associated abstracts and filtered these entities using the list of curated entries for that article. This process created a large quantity of imperfect training data in a very short time. Using a process similar to Collier, an HMM was trained and used to extract relevant terminology. The resulting *F*-score was 75%, quite comparable to Collier's report. This method has the advantage of being rapidly transferable to new domains wherever similar resources exist.

Shen et al. [54] used feature selection to identify lexical features that can capture the characteristics of a biomedical domain. Using HMM, they determined the additive benefit of (1) simple deterministic features such as capitalization and digitalization, (2) morphological features such as prefix/suffix, and (3) part-of-speech features, and compared these features alone as compared to adding (4) semantic trigger features such as head nouns and special verbs. Head noun trigger features enable classification of n-grams. For example the *n*-gram "activated human B-cells" would be classified as "B-cells". Similarly, "special verb trigger" features were verbs that proved useful in biomedical documents for extracting interactions between entities such as "bind" and "inhibit". The GENIA Corpus was used as the training and evaluation corpus. The GENIA corpus (Ver. 1.1) [93] is a human-annotated corpus of 670 biomedical journal abstracts taken from the MEDLINE database, which includes annotations of 24 biomedical classes by domain experts. The overall *F*-score was 66.1% which is 8% higher than Kazama's work [55], which used the identical data set. Simple deterministic features only achieved an *F*-score of 29.4%. Addition of morphological features increased the *F*-score to 31.8%. Addition of POS features provided the largest boost, increasing the *F*-score to 54.3%. Head nouns provide an additional improvement, leading to an *F*-score of 63.0%. But special verb trigger features did not increase the *F*-score at all. They speculated that past and present participles of some special verbs often play the adjective-like roles inside the biomedical terms and may have influenced the classification. For example, in the phrase "IL10 inhibited lymphocytes", the term "inhibited" is a past participle, linking two terms which are not taxonomically related. This may limit the accuracy of this method for taxonomic classification, but suggests that other kinds of ontological relationships could be derived using this method.

Support vector machine (SVM) has also been utilized for biomedical named entity extraction (NER) and subsequent classification. Both Kazama [55] and Yamamoto [56] used the GENIA corpus as training data. Kazama formulated the named entity rec-

ognition as a classification task, representing each word with its context as three simple features (termed "*BIO*") to facilitate the SVM training. *B* indicates that the word is the first word in the named entity, *I* indicates that the word is in another position in the named entity, and *O* indicates that the word is not a part of the named entity. *B* and *I* can be further differentiated by the named entity class annotated within GENIA. Thus, there can be a total of 49 (2 N + 1) classes when the BIO representation is used. For example in the sentence fragment "Number of **glucocorticoid receptors** in **lymphocytes** and ...", where "glucocorticoid receptors" has been human annotated as a member of the class PROTEIN and "lymphocytes" has been human annotated as a member of the class CELL-TYPE, the sentence fragment can be represented as:

Number	of	glucocorticoid	receptors	in	lymphocytes	and	...
O	*O*	B_{protein}	I_{protein}	*O*	$B_{\text{cell-type}}$	*O*	

Because the GENIA corpus has a skewed distribution of classes with the majority of words belonging to the O class, Kazama used a splitting technique to subclass all words in the O class based on POS information. This technique not only made training feasible but also had the added benefit of improving accuracy, because in NER we need to distinguish between nouns in the named entities and nouns in ordinal noun phrases which do not participate in named entities. Kazama achieved an average *F*-score is 54.4% using these techniques.

Yamamoto [56] explored the use of morphological analysis as preprocessing for protein name annotation using SVM. He noted that Kazama's work ignored the fact that biomedical entities have boundary ambiguities that are unlike general English. For example, in general English it may be assumed that the space character is a token delimiter. In contrast, named entities in biomedical domains are often compound nouns, where the space character is not a token delimiter. Consequently, simple tokenization and POS tagging developed for general English may not be adequate for biomedical domains. They proposed a new morphological analysis method that identifies protein names by chunking based on morphemes (the smallest units determined by morphological analysis) as well as word features. This method can avoid the under-segmentation problem that often exists with traditional word chunking. Thus, if a named entity appeared as a substring of a noun phrase, chunking based on noun phrase only would fail to identify it because of coarse segmentation. For example, for the noun phrase "SLP-76-associated substrate", use of a traditional chunking method would only tokenize "SLP-76-associated substrate". In contrast, Yamamoto's morpheme-based chunking method would tokenize both "SLP-76" and "SLP-76-associated substrate". Using the GENIA corpus 3.01, they achieved an *F*-score of 70% for protein names and an *F*-score of 75% for protein names including molecules, families and domains. They suggest that this preprocessing method can be easily adapted to any biomedical domain and improve language processing.

Another machine learning algorithm, Conditional Random Fields (CRFs) model has become popular for term extraction due to their advantages over Hidden Markov Models (HMMs) and Maximum Entropy Markov Models (MEMMs) [94]. Like HMMs and MEMMs, CRFs are discriminative probabilistic models that have been applied to a wide range of problems in text and speech processing. However, CRFs permit relaxed independence assumptions about the random variables and use undirected graphic representations that avoid bias toward states with fewer successor states, the major shortfall of HMMs and MEMMs. For example, Chanlekha and Collier [69] used a CRFs based NER module to learn new concepts of a specific semantic type, namely the spatial information of an event. They treated spatial terms as attributes to each event, (the predicate that describes the states or circumstances in which something changes or holds true), and tried to identify the spatial location of an event based on three sets of features about the event. First, they studied what kind of textual features that people often used to perceive the place where an event in a news report occurred and found 11 of them such as, "Location of the subject" and "Location of the object", which can be used to train the CRFs model. For example, the location of the subject can often indicate the location where an event occurred. In this sentence, "Head of South Halmahera district health office, Dr. Abdurrahman Yusuf confirmed the spread of diarrhea and malaria in the villages". The "South Halmahera district" indicates the location of the subject "Dr. Abdurrahman Yusuf", and it is a clue for where the event, "confirmed the spread of diarrhea and malaria" occurred. Second, they discovered that the type of event could also be utilized as a beneficial feature for spatial term extraction. Using an automatic classifier that they developed, Chanlekha and Collier categorized the events into three groups: spatially locatable event, generic informational event, and hypothetical event. Third, they incorporated the subject type: disease, pathogen, symptom, government or medical officers, person, organization, and location into the feature set. For evaluation, they compared the CRFs with two other methods, (a simple heuristic approach and a probabilistic based approach), on spatial term recognition from a set of 100 manually annotated outbreak news articles from the BioCaster corpus. Using *n*-fold cross validation, they found that CRFs approach achieved the highest performance, (precision 86.3%, recall 84.7%, and *F*-score 85.5%), when compared with a probabilistic approach (precision 69%, recall 74.3%, and *F*-score 71.6%) and simple heuristic approach (precision 52.8%, recall 51.2%, and *F*-score 52%).

3.2. Extraction of taxonomic relationship

Extraction of taxonomic relationships has been extensively studied, using both symbolic and statistical methods.

3.2.1. Symbolic methods

One of the earliest attempts to derive relationships from text corpora was described by Hearst [37], who used lexico-syntactic patterns for semantic knowledge extraction. She hypothesized that linguistic regularities such as LSPs within a corpus can permit identification of the syntactic relationship between terms of interest, and therefore can be used for semantic knowledge acquisition. To prove this hypothesis, Hearst searched for a set of pre-defined lexico-syntactic patterns that indicated general relationships such as hyponym/hypernym in Grolier's American Academic Encyclopedia text. Out of 8.6 million words in the encyclopedia, there were 7067 sentences that contain the pattern 'such as' from which 330 unique relationships were identified and 152 relationships involved unmodified nouns for both hypernym and hyponym, comprising a total of 226 unique words. Using WordNet as a validation resource, she found 180 of these 226 words were present in the WordNet hierarchy, suggesting that these linguistic rules extract meaningful information. She concluded that the LSP matching method could be an effective approach for finding semantically related phrases in a corpus because (a) the method does not require an extensive knowledge base; (b) a single, specially-expressed instance of a relationship is all that is required to extract meaningful information; and (c) the method can be applied to a wide range of texts. She acknowledged low recall as an inherent problem with this method.

Other researchers have applied the LSP matching approach to other domains and investigated methods to increase recall and precision of the LSP approach. Caraballo [70] addressed the low recall problem by applying noun coordination information to the LSP method. Coordination is a complex syntactic structure that links together two or more elements, known as conjuncts or conjoins. The conjuncts generally have similar grammatical features (e.g.

syntactic category, semantic function). He assumed that nouns in a coordination structure, such as conjunction and appositives, are generally related as has been discussed previously by Riloff and Shepherd [95] and Roark and Charniak [96]. For example in the sentence "Sugar, honey, nutmeg, and cloves can increase the flavor of a dish" nutmeg and cloves share a conjunction structure, and are therefore considered to be semantically similar. If "spice" is known to be a hypernym to "nutmeg," then from the sentence above, it can be inferred that "spice" is also a hypernym to "cloves." This linguistic structure can be observed often in biomedical corpora, for example in the sentence: "In the ovine brain, GnRH neurons do not contain type II glucocorticoid (GR), progesterone (PR), or α estrogen (ERα) receptors". Thus, if α estrogen receptor (ERα) is a steroid receptor in the ontology, we can define GR and PR as steroid receptors also.

Cederberg and Widdows [71] described two additional methods that can be added to the extraction process to increase recall and precision. In the first method, they used a graph-based model of noun-noun similarity learned automatically from coordination structures. This method is very similar to Caraballo's method using coordination information. But in contrast to Caraballo's hierarchy-building method, Cedeberg used an alternative graphic-based clustering method developed by Widdows [97] in which nouns are represented as nodes and noun-noun relationships are represented as edges. In Cederberg's graph, the edges between two nouns are connected if they appear in a coordination structure. The algorithm extracts similar words when a seed word is provided by the user, where the seed word is normally a known hyponym of one category. For example, if "clove" is the seed word and is a hyponym of "spice", then all the words that appear in the coordination structure will be hyponyms of "spice" as well. This method obtained additional hypernym–hyponym pairs extracted by LSPs and improved recall 5-fold.

In the second method, Cederberg and Widdows used latent semantic analysis [98] [99] to filter the LSP-extracted hyponyms. Latent semantic analysis is a statistical method that can measure the similarity of two terms based on the context in which they appear. Each term's context is represented by a vector of words that co-occur most frequently with the target term. Similarity between two terms was calculated using the cosine of the angle between the two vectors. A hyponym and its hypernym extracted with the LSP matching method should be very similar. Therefore, by establishing a threshold, term pairs with low scores can be filtered and excluded from further consideration. Using this method, they increased precision of LSP matching from 40% to 58%.

Within the biomedical domain, Fiszman et al. [72] have shown that the Hearst lexico-syntactic patterns can be used for hypernymic propositions to improve the overall accuracy of the SemRep semantic processor developed by Rindflesch and Fiszman [100,101]. SemRep uses syntactic analysis and structured domain knowledge such as the SPECIALIST lexicon and UMLS Semantic Network to capture semantic associations in free-text biomedical documents such as MEDLINE. For example, given a sentence "Alfuzosin is effective in the treatment of benign prostatic hyperplasia", SemRep produces the semantic predication: Alfuzosin-TREAT-Prostactic Hypertrophy, Benign. SemSpec is an extension to SemRep that utilizes LSPs such as appositive structures and Hearst patterns (e.g. "including", "such as" and "especially") to identify hypernymic propositions. Once a hypernymic proposition is established, the more specific term can replace the more general terms in a semantic association that has been captured by SemRep. For example, for a sentence "market authorization has been granted in France for **pilocrapine**, an old **parasympathomimetric agent**, in the treatment of xerostomia "SemRep produces "Parasympathomimetric Agents-TREATS-Xerostomia" and captures the hypernymic position "Pilocrapine-ISA-Parasymphomimetic Agents". From this, a more accurate semantic association "Pilocrapine-TREATS-Xerostomia" can be inferred. Using a manually tagged set of 340 sentences from MEDLINE citations and limited to the UMLS Semantic Network predicate TREATMENT, they found that SemSpec increased SemRep's recall by 7% (39–46%) and precision by 1% (77–78%).

The LSP matching method can be further improved by using machine learning methods to learn LSP patterns. Snow [73] represented the Hearst patterns using a dependency parse tree and found all features along the path for each LSP. These features were used to train a classifier. Snow not only re-discovered the Hearst's patterns, but also identified several new patterns. Riloff [74] developed the Autoslog-TS system, which uses a bootstrapping method for generating LSPs from untagged text. This system is an extension of her earlier Autoslog work [102] and has been further extended in Thelen and Riloff's [103] Basilisk system for semantic lexicon extraction. The input for Autoslog-TS was a text corpus and a set of seed words that belonged to six semantic categories (building, event, human, location, time and weapon). The seed words were generated by sorting all words in the corpus based on frequency, and then manually assigning high frequency words to a category. For example, "bomb", "dynamite", "guns", "explosives", and "rifles" are seed words for "weapon". Seed words were then used to extract contiguous lexico-syntactic patterns, and then the resulting patterns were ranked based on their tendency to extract known category terms. The top patterns were used to extract other terms. Extracted terms were scored, and those with high scores were added into the semantic lexicon. Using a bootstrapping method, this process was then repeated multiple times. The MUC-4 corpus was used to evaluate performance of both Autoslog and Autoslog-TS pattern extraction for aiding semantic information extraction. Autoslog achieved 62% recall and 27% precision, while Autoslog-TS achieved 53% recall and 30% precision. The difference between Autoslog and Autoslog-TS is that Autoslog-TS creates a pattern dictionary with un-annotated training text, whereas Autoslog uses annotated text and a set of heuristic rules. This method has some specific advantages in biomedicine, because of the breadth of resources available for obtaining seed words for a particular semantic category. For example, "ATP", "kinase", "gene transcription", and "binding site" are seed words for "cell activation", which can be obtained from the UMLS or existing biomedical ontologies. As an example, Markó and Hahn [104] have developed a methodology for automatic acquisition and argumentation of a multilingual medical subword thesaurus using seed terms from the UMLS Methathesaurus.

Another linguistic technique for relationship extraction uses compound noun information. For example, Velardi [75] and Cimiano [76] used the following head matching heuristic for hyponym term discovery: *IF term A and term B head nouns are the same and term A has an additional modifier THEN term A is a hyponym of term B.* Using a tourism domain corpus, Velardi achieved 82% precision while Cimiano achieved 50% precision. However, the precisions obtained from different studies are not directly comparable due to the different corpora used.

Rinaldi [77] further expanded Hamon's work, by using Hamon's method to extract all the synsets for each concept and adding the following simple heuristic to organize these synsets into a taxonomic hierarchy: *IF term A is composed of more individual terms than term B, THEN term A is a hyponym of term B.* A manual expert evaluation found 99% accuracy for synonym discovery and 100% accuracy for hyponym links. Morin et al. [78] tried to add a hypernym relationships by mapping one word terms to multi-word terms. For example, given a link between "fruit" and "apple," a relationship between the multi-word terms "fruit juice" and "apple juice" can be added. Similar examples are frequent in biomedical domains. For example, given a relationship between "nucleotide" and "ATP",

a relationship between the multi-word terms "nucleotide transport" and "ATP transport" can be added. Morin et al. based their work on several heuristics: *IF (1) two multi-terms share the same head noun (juice); and (2) the substituted words have the same grammatical function (modifiers); and (3) the substituted words are semantically similar ("apple" and "fruit"), THEN the two terms are related*. For the third clause of the heuristic, semantic information would come from an existing semantic resource such as an ontology. For their knowledge resource, Morin et al. used the Agrovoc Thesaurus, a multilingual thesaurus in the agriculture domain managed by the Food and Agriculture Organization of the United Nations. This method could potentially be very effective in the medical domain because multi-word terms are quite common. For example, terms like "diabetes mellitus" and "insulin-dependent diabetes mellitus" are likely to express taxonomic relationships.

Bodenreider and colleagues [79] explored how to use modifier information to establish groups of similar terms. A group of compound nouns were collected from MEDLINE citations. From these terms, they tried to discover concept candidates for the UMLS Methathesaurus by comparing terms extracted from MEDLINE to current UMLS concepts. They parsed each component noun into a modifier and head noun using an underspecified syntactic analysis [100] and the SPECIALIST Lexicon. The component noun became a concept candidate *if: (1) the head noun of the component noun is found in the Methathesaurus and (2) concepts existing in the Methathesaurus have the same modifier*. The concept candidate was incorporated into the Methathesaurus based on the head noun's position in the hierarchy. From three million randomly selected MEDLINE component nouns, 125,464 of them were captured as concept candidates for Methathesaurus. Evaluation of a sample of randomly selected concept candidates determined how well these candidates can be incorporated into the Methathesaurus using head noun matching. The authors defined three levels of relevance: The highest level, "relevant", was used for cases where the addition of the candidate to the terminology was relevant even if there was a more specific concept available. The intermediate level, "less relevant", was used for cases where the parent selected for the candidate is too general to be informative. The lowest level, "not relevant", was used for cases where the parent selected for the concept is irrelevant. From 1000 randomly selected candidates, 834 were classified as "relevant", 28 were classified as "less relevant" and 138 were classified as "irrelevant".

Investigating an alternative approach to heuristics, Ryu [80] explored a mathematical method for determining hierarchical position using 'specificity' as defined in the field of information theory [105], where specificity of a term is a measure of the quantity of domain-specific information contained in the term. Therefore, the higher the specificity of the term, the more specific the information it contains (further details regarding this measure are discussed in Section 3.2). A weighting scheme was used to exclude terms that frequently appear as modifiers but provide no additional information. The taxonomic position of the term was then determined based on the specificity. For example, "insulin-dependent diabetes mellitus" had a higher specificity and thus should be positioned as a child of "diabetes mellitus". Using a flat collection of terms obtained from a sub-tree of MeSH and a set of journal abstracts retrieved from the MEDLINE database, the authors generated a hierarchy for the MeSH terms, and compared it to the MeSH hierarchy. The precision for ontological hierarchy placements was increased from 69% to 82% when compared against a word frequency baseline method.

3.2.2. Statistical methods

Both clustering and machine learning methods have also been applied to the extraction of relationships, albeit less frequently and with less success than extraction of concepts. Alfonseca and Manandhar [58] followed Agirre's [48] topic signature technique. With a top-down search, starting with the most general concept in the hierarchy, the new concept was added to the existing concept whose topic signature was the closest to that of the new concept's. Several experiments were conducted with seven general target words. The task was to place these words into the right category in the ontology. The best result was 86% accuracy. They also concluded that it was better, for this task, to consider a smaller context of highly related words to build the signature rather than a larger context that included more words.

Another group led by Witschel [81] extended a decision tree model for taxonomy enrichment. They first identified potential new concepts using a combination of statistical and linguistic methods [106] termed "semantic description" based on co-occurrence within German language texts (such as newspapers, fiction, etc.). Witschel's 'semantic description' is similar to Alfonseca's 'distributional signature' [57]. They evaluated their method using a general-German language text to enrich a sub-tree of GermaNet (the German equivalent to WordNet). Two measures were computed - accuracy of the decisions (percentage of nodes that were correctly classified as hypernyms) and learning accuracy [60] which takes into consideration the distance of the automated placement from the expected location in the tree. The accuracies for enriching a furniture sub-tree and a building sub-tree were 11–14% respectively which was comparable to Alfoneca's result. The learning accuracy reached 59% which was significantly better than Alfoneca's. Again, the absence of a common reference standard for testing makes it difficult to directly compare these results.

3.3. Extraction of non-taxonomic relationships

Extraction of non-taxonomic relationships, i.e. non-IS-A relationships, has also been studied, and has been considered to be the most difficult ontology learning task. Both symbolic and statistical methods have been employed.

3.3.1. Symbolic methods

The LSP method has been used by Berland [82], Sundblad [83], and Girju [84] for part-whole (meronymic) relationship discovery. Berland combined both the LSP method and statistical methods and used them on a very large corpus. The output of the system was an ordered list of possible parts for a set of six seed "whole" objects. They achieved 55% accuracy.

Nenadić and Ananiadou [85] used three symbolic approaches to discover terms from MEDLINE abstracts: (1) lexico-syntactic pattern based similarity measure (SS) using Hearst patterns, coordination patterns, apposition patterns, and anaphora, (2) a component noun based similarity measure (which they called the lexical similarity measure (LS)), and (3) contextual pattern based similarity measure. The third approach, which was considered novel by the author, learns contextual patterns by discovering significant term features. The procedure is performed as follows and illustrated using our ATP example. First, for each target term, its context constituents are tagged with POS tags and grammatical tags. These tags became the context pattern for the target term. For example, in the phrase "ATP binds heterodimers with ***high affinity***", "high affinity" is the target term, and the left context pattern (CP) is "*V: bind TERM: rxr_heterodimers PREP:with*". Second, all the CPs for each term are collected and a normalized CP-value is calculated in order to measure the importance of the CP. The CP-value is calculated based on the length and the frequency of the pattern. The similarity between two terms based on CP was termed CS (t_1, t_2) and calculated based on the number of common and distinctive CPs of the two terms. Since none of the three similarity measures is sufficient on its own, they introduced a hybrid term similarity

measure called Contextual Lexical Similarity (CLS) which is a linear combination of the three similarity measures with three parameters: $CLS(t_1, t_2) = \alpha\ CS(t_1, t_2) + \beta\ LS(t_1, t_2) + \gamma\ SS(t_1, t_2)$. In the final step, the three parameters (α, β, γ) were adjusted automatically by supervised learning methods. They tested the CLS measure on a corpus of 2008 abstracts retrieved from MEDLINE. Random samples of results were evaluated by a domain expert to see if the two similar terms based on CLS measure were indeed similar. They also used the CLS measure for term clustering and achieved a precision of 70%.

3.3.2. Statistical methods

Kavalec [86] uses a statistical approach, supplemented with some linguistic information to extract non-taxonomic relationships. In this case, the linguistic feature used was based on the assumption that relational information is typically conveyed by verbs at the sentence level. For example, the verb "induce" defines a non-taxonomic (associational) relationship between a gene and a protein. Therefore, he first selected verb v and a pair of concepts that co-occur within a certain window of verb v. Second, the concept–concept–verb triples were ordered by frequency. The highest frequency triples were candidates for relationship labels of the given concept association. The association measure was a simple statistical measure based on a verb and a concept pair conditional frequency (co-occurrence), $P(c_1, c_2|v)$. However, the conditional frequency of a pair of concepts, given a verb, could be high even though there is no relationship between the concepts and the verb. This occurs because a verb may occur separately with each of the concepts at high frequency, even though it has nothing to do with any of the mutual relationships between the two concepts. Therefore, the authors defined an "above expectation" (AE) measure (see Eq. 1 below), which was a measure of the increased frequency when compared to the frequency expected under the assumption of independence of association of each of the concepts with the verb. This measure is very similar to the "interest measure" suggested by Kodratoff [107] for knowledge discovery in text, and also the Church mutual information metric [42].

$$AE(c_1, c_2|v) = \frac{P(c_1, c_2|v)}{P(c_1|v) \cdot P(c_2|v)} \quad (1)$$

The authors performed several experiments to evaluate this approach. In one of the experiments, an ad hoc tourist document collection was used as input for the method. In another experiment, the SemCor corpus that had been semantically tagged with the WordNet senses was used. The results were promising. At AE = 1.5 (1 is equal to expectation value), the recall was 54% and precision was 82% for the tourist corpus (measured against a human annotated reference standard). For the SemCor corpus, expert judges evaluated the output, yielding a precision of 72%. Recall could not be measured.

An alternative statistical approach uses association rule mining methods to extract relationships between concepts [49–51]. This method was first introduced by Agrawal et al. [108] as a technique for market analysis using a large database of transaction data. The rules extracted can be exemplified as "90% of the transactions that purchased bread and butter also purchased milk". The method has been adapted to mine domain text for concept relationships. The advantage of this method is it does not require deep textual analysis. However, it tends to generate a large number of association rules. Statistical indices such as support and confidence are then used to select the most meaningful and significant rules. Although the method does not distinguish among type of relationships, it could easily be used as a starting point for human curation.

Gulla [49] evaluated and compared this method with traditional similarity measure methods that utilize vector space models. The output was judged by four human experts by separating extracted relationships into three categories: "not related", "related" and "highly related". The results shown that more than half of the relationships found by association rule methods were also identified by the similarity measure method. However, the distribution of mined rules was different using these two methods. A further experiment combining the methods produced much better results. They concluded that these two methods may be complementary when combined for relationship extraction. Cherfi's work [50] focused on investigating how the characteristics of several statistical indices such as support, confidence, interest, conviction and novelty influence the performance of association rule mining and how a combination of different indices ensure that a subset of valid rules will be extracted.

In the biomedical domain, Bodenreider et al. [51] evaluated and compared the association rule method (ARM) with two other statistical methods that use similarity measures: the vector space model (VSM) and co-occurrence information (COC), for identifying associations of GO terms between three GO sub-ontologies (molecular function, cellular components, biological processes). They took advantage of several existing databases of human annotations using GO terms that were publicly available. For the VSM method, gene products that associated with the GO term in the databases were used to form a vector and the similarity of two GO terms was calculated as the cosine of the two vectors. For the COC method, the frequencies of co-occurring GO terms in the database was represented as a contingency table (number of gene products annotated with both term A and B, number of gene products annotated with term A only, number of gene products annotated with term B only, number of gene products annotated with neither term A or B), and a chi-square test was used to test the independence of the two GO terms. If the terms were not dependent, they were considered to be associated. For the ARM method, each annotation of gene products with GO terms was treated as a transaction. Association rules were extracted using the Apriori algorithm [109]. They evaluated the validity of the extraction by comparing the overlap between the statistical methods, and by comparing statistical methods to another set of methods that were non-statistical and not based on a document corpus. These non-statistical methods included extracting relationships between GO terms existing in UMLS or MeSH (where the relationship is not also included in GO), and determining lexical relationships based on composition between existing between GO terms (where the relationship is not also included in GO). A total of 7665 associations between GO terms were identified by at least one of the three statistical methods (VSM, COC, and ARM). Among 7665 associations extracted by these statistical methods, 936 (12%) of them were identified by at least two of the three statistical methods and 201 (3%) of them were identified by all three statistical methods. Using the non-statistical methods, 5963 associations were identified. But the authors note that when comparing the relationships extracted by statistical methods to those obtained using the non-statistical methods, only 230 overlapping associations were found. They conclude that multi-method approaches may be necessary to extract a more complete set of relationships.

3.4. De novo generation of ontologies

In contrast to the process of ontology enrichment (which seeks to add or modify existing ontologies), a few researchers have explored the possibility of learning the entire ontology by combining methods for multiple tasks.

Lin [45] explored the distributional pattern of dependency triples as the word context to measure word similarity. Lin's work is very similar to Grefenstette's approach [110] in which dependency triples were treated as features. A dependency triple consists

of two words and the grammatical relationship between them in the input sentence. As an example in our own domain, the triples extracted from the sentence "The patient has a mild headache" would be "(has subj patient), (patient subj-of has), (headache obj-of has), (headache adj-mod mild), (mild adj-mod-of headache), (headache det a), (a det-of headache). The description of a word w consists of the frequency counts of all the dependency triples that matched the pattern (w, *, *). Therefore, the similarity between two words was calculated based on the count of dependency triples for each word. Using this similarity measure, Lin created a thesaurus and evaluated this thesaurus against WordNet and Roget Thesaurus. He found his thesaurus was more similar to WordNet than Roget Thesaurus and using all types of dependency triple was better than using only subject and object triples as Hindle did [46].

Blaschke and Valencia [87] explored the statistical clustering method for building an ontology-like structured knowledge base using the biomolecular literature. They adapted Wilbur's [111] method by clustering the key terms that have been derived from the documents associated with each individual gene. They first retrieved over 6000 gene names associated with *Saccharomyces cerevisiae* from SWISS-PROT and SGD. 63,131 MEDLINE abstracts were obtained with search terms "*saccharomyces*" and/or "*cerevisiae*". Then, they grouped the documents based on each gene name they associated with and created a fingerprint for each group that could describe the specific content of documents. The fingerprint consisted of a list of key-terms (including bi-grams) and the scores for each term. The score was calculated by comparing frequencies between groups of documents. This fingerprint was used to calculate the similarity between two genes, a and b, ($SimScore_{a,b}$) as the sum of the scores for all significant terms i that appear in both fingerprints.

$$SimScore_{a,b} = \frac{\sum(score_i^a + score_i^b)}{2} \quad (2)$$

To construct the ontology, a distance matrix for all pairs of genes was created by calculating the similarity score for each pair of genes. Two genes with the highest score were clustered together and removed from the distance matrix, and the two groups of documents for these two genes were merged. A new fingerprint for the merged documents was created. This process was repeated until none of the clusters shared more significant terms. The final output was a gene tree, which was compared with the hand-curated GO ontology by domain experts and found by them to be compatible. Some relationships in the tree that were not in the GO could be added. They concluded that this automatic clustering method can be utilized as an instrument to assist human expert's ontology building. This approach could be particularly useful for domains experiencing rapid growth. For example, in genomics, many new genes have been discovered as a result of the advances in genomic sequencing. The number of potential relationships among these genes and proteins is quite large and therefore could be amenable to a semi-automated approach.

4. Existing ontology learning systems

In recent years, a number of ontology learning systems have been developed using one or more of the algorithms described above with the goal of reducing the human effort required for ontology development. In this section, we compare eleven state-of-the-art ontology learning systems. Three of these systems were developed primarily for the biomedical domain, and the remaining eight systems were developed for general language or other domains. We examine and compare the elements learned from the text as well as the different approaches employed and different evaluations performed. Table 2 summarizes these comparisons. All eleven systems are able to learn concepts and taxonomic relationships. Additionally, the DOODLE II, HASTI, STRING-IE, Text-To-Onto and Text2Onto systems can also learn non-taxonomic relationships.

ASIUM [112] (Acquisition of Semantic Knowledge Using Machine Learning Methods) is a system developed to acquire ontological knowledge and case frames. The input to the system is a set of domain-specific documents in French that have been syntactically parsed. The system uses clustering methods, based on a two-step process which produces successive aggregations. The first step is conceptualization clustering which is similar to Harris [38], Grefenstette [41] and Peat's [113] work, in which the head words associated with their frequencies of appearance in the text are used to calculate the distances among concepts. Based on the sub-categorization of verbs, the head words that occur with the same verb after the same preposition (or with the same syntactical role) are clustered into the basic cluster. The second step is pyramidal clustering that they adopted from Diday [114], in which the basic clusters are built into the hierarchy of the ontology [115]. This approach is promising, but an evaluation with real cases and real problems has not yet been performed.

DODDLE II [116] is a domain ontology rapid development environment. The inputs to the system are a machine-readable dictionary and domain-specific texts. It supports both the building of taxonomic and non-taxonomic relationships. The taxonomic relationships come from WordNet. The non-taxonomic relationships come from domain-specific text and from analyzing the lexical co-occurrences based on WordSpace [117] which is a multi-dimensional, real-valued vector space representing lexical items according to how semantically close they are. Evaluation was done in the domain of Law with two small-scale case studies. One study used 46 legal terms from Contract for the International Sale of Goods part II (CISG) and the other study used 103 terms that included general terms from the CISG corpus. For taxonomic relationships, the precision was 30%. For non-taxonomic relationships, the precision was 59%.

HASTI [118] is a system that learns concepts, taxonomic and non-taxonomic relationships, and axioms. It is the only system that also learns axioms from text documents (in Persian). HASTI employs a combination of symbolic approaches such as Hearst patterns [37], logic, template, as well as semantic analyses and heuristic approaches. It has two modes for conceptual clustering: automatic and semi-automatic. HASTI requires only a very small kernel of an ontology containing essential meta-knowledge such as primitive concepts, relations and operators for adding, moving, deleting and updating ontological elements. Based on this kernel, the system can learn both lexical and ontological knowledge. The kernel is language neutral and domain independent. Therefore, it can be used to build both general and domain ontologies, essentially from scratch. To prove that the system can be generalized, the authors evaluated HASTI with two test cases. With a text corpus consisting of primary school textbooks and storybooks, the precision was 97% and the recall was 88%. With a text corpus consisting of computer technical reports, the precision was 78% and the recall was 80%.

KnowItAll [119] is an automatic system that extracts facts, concepts, and relationships from the WWW. There are three important differences between this system and other similar systems, First, KnowItAll addresses the scalability issue by using weakly supervised methods and bootstrapping learning techniques. Using a domain-independent set of generic extraction patterns, it induces a set of seed instances, thus overcoming the need for a hand-coded set of training documents which is typically required for these kinds of systems. Second, it uses Turney's PMI-IR methods [88] to assess the probability of extractions using statistics computed by treating the web as a large corpus of text (so called "web-scale

Table 2
Characteristics of nine existing ontology learning systems.

	Input	Language	Ontological elements learned	Degree of automation	Resource	Ontology enrichment or de novo generation	Learning Methods
ASIUM	Free text documents	French	Concepts taxonomic relations	Semi-automated	N/A	DenoVo	Conceptual and hierarchical clustering
DODDLE II	Dictionary domain-specific text documents	English	Concepts taxonomic relations, non-taxonomic relations	Semi-automated	WordNet	Enrichment	Matching and trimming against WordNetfor taxonomic relations, statistical co-occurrence information
HASTI	Free text documents	Persian	Concepts taxonomic relations non-taxonomic relations, axioms	Two modes: semi-automated and fully-automated	N/A	DenoVo	Combination of logical linguistic template, and heuristic
KnowItAll	Web pages	English	Concepts	Automatic	Domain ontology	Enrichment	Combination of linguistic and statistic methods
MEDSYNDIKATE	Medical domain documents	German	Concepts taxonomic relations	Semi-automated	Own general and medical lexicons; Fully lexicalized dependency grammar	Enrichment	Input text is mapped to corresponding text knowledge basestTKB) which represent the text content: Generates concept hypothesis and ranks hypothesis based on quality
OntoLearn	Free text documents	English	Concepts, taxonomic relations	Semi-automated	WordNet: SemCor	Enrichment	Machine learning statistical approach
STRING-IE	Free text documents from PubMed	English	Non-taxonomic relations	Automated	SWISS-PROT Saccharomyces Genome Database	Enrichment	Linguistic and rule based approach
Text-To-Onto Text20nto	Dictionaries databases semi-structured text. Free text documents	German	Concepts taxonomic relations non-taxonomic relations	Semi-automated	Domain ontology (Tourism)	Enrichment	Combination of association rules formal concept analysis and clustering
TIMS	Free text documents	English	Concepts taxonomic relations	Automated	N/A	Enrichment	Automatic term recognition using both linguistic and statistical approach and automatic clustering using average mutual information
WEB- > KB	Web pages	English	Concepts taxonomic relations	Automated	Domain ontology	Enrichment	Statistical and logical

statistics"). This overcomes the problem of maintaining high precision and enables the system to automatically trade recall for precision. Third, it is able to make use of the ample supply of simple sentences on the WWW that are relative easy to process, thus avoiding the extraction of information from more complex and problematic texts. Details of the algorithmic methods [63] were described earlier in Section 3.1.1.

MEDSYNDIKATE [120] is an extension of the SYNDIKATE system. It is the only knowledge acquisition system aimed at acquiring medical knowledge from medical documents (in German). MEDSYNDIKATE enables the transformation of text documents to formal representation structures. The system addresses one of the shortcomings of information extraction systems by providing a parser that is particularly sensitive to the treatment of textual reference relationships as established by various forms of anaphora [121]. It distinguishes between text at the sentence level and the text level. A deeper understanding of textual referential relationships is based on their *centering* mechanism [122]. Additionally, MEDSYNDIKATE initiates a new conceptual learning process (knowledge enrichment) while understanding the text. Domain knowledge and grammatical constructions such as lexico-syntactic patterns in the source document in which the unknown word occurs are used to access the linguistic quality and conceptual evidence. This information is used to rank the concept hypotheses. The most credible hypotheses based on ranking are selected for assimilation into the domain knowledge base. Another technique for concept generation is based on the reuse of available comprehensive knowledge sources such as UMLS. Evaluation of MEDSYNDIKATE was performed on the deep semantic understanding of the input text but not on the concept learning aspect of the system. Although this is a system developed for the medical domain, the German language basis of the system may somewhat limit its transfer to English language documents. Nevertheless, methodologies developed and insights derived from MEDSYNDIKATE are extremely valuable to researchers developing ontology enrichment systems for English language documents in biomedical domains.

OntoLearn [75] is a very sophisticated system, that uses a combination of symbolic and statistical methods. Domain-specific terms are extracted and related to corresponding concepts in a general purpose ontology and relationships between the concepts are examined. First, statistical comparative analysis is done on the target domains and the contrasting corpora to identify terminology that is used in the target domains but not the contrasting corpora. Second, lexical knowledge of WordNet is used to interpret the semantic meaning of the terms. OntoLearn then organizes the concepts based on taxonomic and non-taxonomic relationships into a forest using WordNet and a rule-based inductive learning method. Finally, it integrates the domain concept forest with WordNet to create a pruned and specialized view of the domain ontology. The validation of the process is performed by an expert.

The system has been evaluated by two human judges, across a variety of ontology learning algorithms. The evaluation results are encouraging. With different domains (art, tourism, economy and computer network), they achieved recall ranging from 46% to 96% and precision ranging from 65% to 97%.

STRING-IE [123] is a system designed to extract non-taxonomic relationships between concepts in biomedical domain using symbolic features and rules (heuristic). More specifically, it extracts regulation of gene expression and (de-)phosphorylation related to yeast *S. cerevisiae*. Although the language rules they created are specific for *S. cerevisiae* organism, they have tested their algorithm on three other organisms (*Escherichia coli*, *Bacillus subtilis* and *Mus musculus*) and achieved equally good results. Therefore, they believe the method is generalizable. The input to the system is a set of abstracts and full text papers from PubMed Central retrieved with terms '*Saccharomyces cerevisiae*', '*S. cerevisiae*', '*Baker's yeast', 'Brewer's yeast'* and '*Budding yeast'*. First, the documents were POS tagged and a name entity recognition was used to identify names of genes and proteins. The NER module uses syntactic-semantic chunking. For example, the text "the ArcB senory kinase in Escherichia coli" would be chunked as "[$_{nx\_kinase}$ [$_{dt}$ the] [$_{nnpg}$ ArcB] [$_{jj}$ senory] [$_{kinase}$ kinase] [$_{in}$ in] [$_{org}$ Escherichia coli]]. The label $_{nx\_kinase}$ indicates this is a noun chunk (*nx*) semantically denoting a *kinase*. After NER, two types of relationships were extracted using heuristics to identify verbs related to these relationships as well as other symbolic features such as the pattern "*x* but not *y*" and pre-defined information about linguistic restriction. They also created a set of rules over groups of verbs and relational nouns, triggered by key words related to regulation of gene expression, such as "phosphorylate", "induce", "decrease", "regulate" and "mediate". For evaluation, they used one million PubMed abstracts that related to the organisms above. A total of 3319 regulatory network and phosphorylation relations were extracted. The accuracy of the extraction was 83–90% for regulation of gene expression and 86–95% for phosphorylation.

Text-To-Onto [124] is a semi-automatic ontology learning system. The system employs a shallow parser (in German) to pre-process text documents coming from the WWW. The advantage of this system is that it has a built-in algorithm library that supports several distinct ontological engineering tasks. The library includes several algorithms for ontology extraction and several algorithms for ontology maintenance such as ontology pruning and refinement. It gives the user the ability to pick extraction and maintenance algorithms for various inputs and tasks. For ontology concept and concept relationship extraction, Text-To-Onto utilizes a combination of statistical methods such as Srikant's [125] generalized association rule discovery and symbolic methods such as Hearst's lexico-syntactic pattern method. Details of extraction algorithms are described in other manuscripts [126–129]. Later, these researchers developed Text2Onto [130] which was distinguished from the earlier system in three important ways. First, they represented the learned knowledge at a meta-level in the form of instantiated model primitives, which they termed the Probabilistic Ontology Model (POM). In this way, learned knowledge remained independent of a concrete target language and could be translated into any knowledge representation formalism (e.g. RDFS, OWL, *F*-Logic). Second, to facilitate user interaction, they used the POM to calculate a confidence for each new learned object. Users could thus filter the POM, selecting only a number of relevant instances of modeling primitives that fit their interests. Third, they explicitly track the changes to the ontology since the last change in the document collection so that users can trace the evolution of the ontology over time as new documents are processed. An obvious benefit is that there is no longer the need to process the entire document collection when additional documents are added later. But, such transparency into the working of the system over time could also enable greater human supervision of the enrichment process.

Both taxonomic relationship discovery using Hearst pattern match method and non-taxonomic relationship discovery using Srikant's [125] generalized association rule discovery method were evaluated in a tourism domain. For taxonomic relationship (IS-A) discovery, they achieved 76% accuracy. For non-taxonomic relationship discovery, they manually developed a small ontology with 284 concepts and 88 non-taxonomic relationships as the gold standard. As the traditional evaluation metrics – (precision and recall), cannot measure the real quality of automatic relationship discovery if the relationships are of varying degrees of accuracy, they defined four categories of relationship matches against the gold standard as "utterly wrong", "rather bad", "near miss", and "direct hit". Then, they defined a new metric called Generic Relation Learning Accuracy (RLA) to measure the average accuracy of an instance of a relationship discovered against the best counterpart from the gold standard. The best RLA was 67% when experimenting with different parameters (support and confidence).

TIMS (Tag Information Management System) [131] is a terminology-based knowledge acquisition and integration system in the domain of molecular-biology. The system is very comprehensive and can support ontology population using automatic term recognition and clustering, knowledge integration and management using XML-data management technology, as well as information retrieval. For knowledge acquisition, TIMS used automatic term recognition (ATR) and automatic term clustering (ATC) modules. The ATR module is based on the C/NC-value method [132] which uses both symbolic information, such as POS tag, and statistical information, such as frequency of occurrence of the term. The C/NC method is specifically adapted to multi-word term recognition. The ATC module is based on Ushioda's AMI (Average Mutual Information) hierarchical clustering method [133] and is built on the C/NC results. The output of ATC is a dendrogram of hierarchical term clusters. Using a NACSIS AI-domain corpus and a set of MEDLINE abstracts, preliminary evaluation of ATR showed precision from 93% to 98% for the top 100 terms.

Focusing on the vast quantity of information available on the WWW, WEB → KB [134] is an ontology learning system that uses a machine learning approach for trainable information extraction. The system takes two inputs: (1) a knowledge base consisting of ontology defined classes and relationships and (2) training examples from the Web that describe instances of these classes and relationships. Based on these inputs, the system determines general procedures capable of extracting additional instances of these classes and rules to extract new instances, rules to classify pages and rules to recognize relationships among several pages. WEB → KB uses mainly statistical, machine-learning approaches to accomplish these tasks. For evaluation, the authors attempted to learn information about faculty, student, course and departments from Web pages, creating an organizational knowledge base. The average accuracy was over 70%, at a coverage level of approximately 30%. They also explored and compared a variety of learning methods, including statistical bag-of-words classifiers, first-order rule learner, and multi-strategy learning methods. They found more complex methods such as first-order rule learning tended to have better accuracy than the simple bag-of-word classifier, at the expense of lower coverage.

5. Conclusions and implications

Previous research has demonstrated that methodologies developed in the fields of Natural Language Processing, Information Retrieval, Information Extraction, and Artificial Intelligence can be utilized for ontology enrichment to alleviate the knowledge acqui-

sition bottleneck. However, there are many issues that must be addressed before we can fully realize the potential benefits of these methods for fully automated or even semi-automatic ontology enrichment in biomedical domains.

Based on our review of the literature, we believe that current methods can be effectively applied to ontology learning in biomedical domains, although some methods may be more useful than others due to the constraints of medical and biological language. Many characteristics of the biomedical domain make it particularly appealing to attempt ontology enrichment using these methods. First, many linguistic features utilized by the various linguistic approaches are quite prevalent in medical and biological text. For example, compound nouns are common in the biomedical domain, because many biomedical terms are composed by adding additional modifiers to the existing terms. A number of researchers have explored this phenomenon in detail [135–137] especially because of its implication for post-coordination and compositional models [137]. Methods that utilize such component information could be effective for hyponym placement [78] [62]. Second, our field has well-developed knowledge and lexical resources such as existing ontologies/terminologies, domain-specific corpora, and general dictionaries that are necessary for knowledge extraction. WordNet provides an important resource for ontology learning of general English domains [48] and could be utilized in ontology learning in biomedical domains. Combined approaches that leverage both WordNet and biomedical ontologies and vocabularies could be particularly interesting. With wide recognition of the importance of sound and complete ontologies in the field of biomedical informatics, endeavors such as the NCI's Enterprise Vocabulary Services, Open Biomedical Ontologies Consortium, and the Gene Ontologies Consortium provide ample opportunities to explore the benefit of enrichment of existing biomedical ontologies.

However, many of these techniques have never been tested and evaluated in biomedical domains. Almost all systems built for ontology knowledge learning and extractions have been developed specifically for domains other than biomedical domain, and often in languages other than English. There are significant barriers to overcome in immediately translating previous research into the biomedical domain.

First, biomedical language is very different than these other domains [138–141]. Many general English based algorithms may not be effective when they are applied for more specific sublanguages. For example, authors of this manuscript have found that some of the simple LSPs, such as Hearst patterns, have low recall in clinical documents, which limits the effectiveness of LSP method for ontology learning using clinical documents [142]. This problem can be alleviated by discovering domain-specific patterns from domain corpus using the pattern learning approach [74,143,73].

Second, sources of biomedical text, such as clinical and biomedical texts, also differ in their characteristics. For example, many clinical reports are structured in such a way that the header or sections provide context which must be used to make inferences regarding further content. For example, in pathology reports we often see text such as "PROSTATECTOMY: Adenocarcinoma", and must infer the origin of the disease from our knowledge regarding the procedure. Few algorithms have specifically addressed the issues related to section segmentation and inference. In general, investigators in this area would benefit by systematically testing and extending existing approaches that can best explore the characteristics of biomedical and clinical text, and directly comparing performance of these methods on biomedical text. The performance of existing methods is likely to vary by domain and task. The authors of this review paper are currently working on an open source development project to make ontology learning methods more widely available to biomedical researchers. The Ontology Development and Information Extraction (ODIE) toolkit [144] is being developed in collaboration with the National Center for Biomedical Ontology (NCBO). As part of this project, we are currently evaluating a number of these algorithms for use in biomedical domains, particularly applied to clinical reports.

Third, a significant challenge for ontology enrichment is the lack of systematic evaluation methods and reference standards. Furthermore researchers working in the same area may be evaluating different aspects of enrichment, and thus cannot be compared. Only a few researchers have dedicated significant work to developing appropriate evaluation methods. The OntoClean methodology [145] developed by Guarino's group describes a set of rules that can be applied systematically to taxonomies to remove the erroneous subclass (in is-a relationship). This may be useful for ontology pruning and refinement. Another group led by Faatz and Steinmetz [146] studied an evaluation framework for ontology enrichment. They described a quality measurement framework for ontology enrichment methods with relevance and overlap heuristics. More research is needed in this area to develop robust performance metrics, and to move the field towards more standardized approaches permitting meta-analysis.

One clear conclusion that we draw from this literature review is that fully automated acquisition of ontology by machines is not likely in the near future. On the one hand, symbolic methods suffer the limitation of coverage and applicability due to the requirement of manual acquisition and codification of lexical knowledge for each domain. On the other hand, statistical methods such as methods based on word co-occurrence information, in general, cannot provide linguistic insight on their own. Therefore, a human expert is required to make sense of the results. As such, a much more practical approach is to develop semi-automatic ontology learning that includes human intervention. With this goal in mind, a perfect or optimal ontology learning method may not be crucial. Relatively simple methods that are appropriately integrated into practice may provide value at a relatively low cost. Systems that suggest potential concepts and relationships could also use information about the curator's judgments to further refine future suggestions, using a bootstrap method. Achieving this goal may require just as much work in optimizing the human–computer interaction as it does in developing algorithms for extracting potential concepts and relationships. Thoughtful integration into existing ontology development workflow is likely to be the key.

In summary, research in the area of ontology learning has focused on non-biomedical domain but has significant potential for enhancing existing biomedical ontologies and reducing the knowledge acquisition bottleneck. We propose that further work to test and extend existing algorithms on biomedical text, integrate them into ontology development workflow, and develop methods for sound evaluation provide the foundation for the development of novel systems that ease the arduous task of developing biomedical ontologies.

Acknowledgments

We wish to thank Wendy Chapman, PhD at University of Pittsburgh, Guergana K. Savova, PhD at Mayo Clinic and Daniel Rubin, MD, MS at Stanford University for valuable comments and feedback about this paper. We thank Karma Lisa Edwards and Lucy Cafeo of the University of Pittsburgh for expert editorial assistance. We thank the two anonymous reviewers of this whose insightful critiques and suggestions have helped us improve the quality and completeness of this review. This work was supported by NIH Grant RO1 CA 127979.

References

[1] de Keizer NF, Abu-Hanna A. Understanding terminological systems II: terminology and typology. Methods Inf Med 2000;39:22–9.
[2] de Keizer NF, Abu-Hanna A, Zwetsloot-Schonl JHM. Understanding terminological systems I: terminology and typology. Methods Inf Med 2000;39:16–21.
[3] Cornet R, De Keizer NF, Abu-Hanna A. A framework for characterizing terminological systems. Methods Inf Med 2006;45:253–66.
[4] William WC, Sunita S. Exploiting dictionaries in named entity extraction: combining semi-Markov extraction processes and data integration methods In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, WA, USA: ACM; 2004.
[5] Navigli R, Velardi P. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. IEEE Trans Pattern Anal Mach Intel (PAMI) 2005;27:1075–86.
[6] Poesio M, Vieira R, Teufel S. Resolving bridging references in unrestricted text, Proceedings of the ACL Workshop on Operational Factors in Robust Anaphora Resolution, 1997, p. 1–6.
[7] Soon WM, Ng HT, Lim DCY. A machine learning approach to coreference resolution of noun phrases. Comput Linguist 2001;27:521–44.
[8] Ng V, Cardie C. Improving machine learning approaches to coreference resolution. In: Proceedings of the 40th Annual Meeting of the ACL. Philadelphia, Pennsylvania: ACL; 2001.
[9] Friedman C, Borlawsky T, Shagina L, Xing HR, Lussier YA. Bio-ontology and text: bridging the modeling gap. Bioinformatics 2006;22:2421–9.
[10] Liang T, Lin Y-H. Anaphora resolution for biomedical literature by exploiting multiple resources. In: Second International Joint Conference on Natural Language Processing, 2005.
[11] Gomez F. An algorithm for aspects of semantic interpretation using an enhanced WordNet. In: Second meeting of the North American Chapter of the ACL on Language Technologies ACL, Pittsburgh, Pennsylvania, 2001.
[12] Gomez-Perez A, Manzano-Macho D. An overview of method and tools for ontology learning from texts. Knowledge Eng Rev 2005;19:187–212.
[13] Girju R, Moldovan DI. Knowledge acquisition for question answering. In: Proceedings of the Fourteenth International Florida Artificial Intelligence Research Society Conference, AAAI Press, 2001.
[14] Bodenreider O, Stevens R. Bio-ontologies: current trends and future directions. Brief Bioinform 2006;7:256–74.
[15] Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, et al. ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res 2008;36:D344–50.
[16] Gruber TR. A translation approach to potable ontology specifications. Knowledge Acquisition 1993;5:199–220.
[17] Smith B, Kusnierczyk W, Schober D, Ceusters W. Towards a reference terminology for ontology research and development in the biomedical domain, in: Bodenreider O, editor. In: The Second International Workshop on Formal Biomedical Knowledge Representation: "Biomedical Ontology in Action" (KR-MED 2006), 2006.
[18] Cimino JJ. In defense of the Desiderata. J Biomed Inform 2006;39:299–306.
[19] Smith B. From concepts to clinical reality: an essay on the benchmarking of biomedical terminologies. J Biomed Inform 2006;39:288–98.
[20] National Institutes of Health. Research Portfolio Online Reporting Tools (RePORT), 2010. Available from: http://projectreporter.nih.gov/project_info_history.cfm?aid=7941562&icde=2611544.
[21] BioInform. Stanford's Mark Musen on the New National Center for Biomedical Ontology, 2005. Available from: http://www.genomeweb.com/informatics/stanford-s-mark-musen-new-national-center-biomedical-ontology.
[22] Du L. DUMC gets $1.25M for ontology, The Chronicle, 2009.
[23] National Science Foundation. The Hymenoptera Ontology: Part of a Transformation in Systematic and Genome Science, 2009. Available from: http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0850223.
[24] United States National Library of Medicine, FAQs: SNOMED CT® in the UMLS®, 2003. Available from: http://www.nlm.nih.gov/research/umls/Snomed/snomed_faq.html.
[25] United States National Library of Medicine. SNOMED Clinical Terms® To Be Added To UMLS® Metathesaurus®, 2003. Available from: http://www.nlm.nih.gov/research/umls/Snomed/snomed_announcement.html.
[26] Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res 2009;37:W170–173.
[27] Tudorache T, Noy NF, Tu SW, Musen MA. Supporting collaborative ontology development in Protege. In: Seventh International Semantic Web Conference, Karlsruhe, Germany, 2008.
[28] Campbell KE, Cohn SP, Chute CG, Shortliffe EH, Rennels G. Scalable methodologies for distributed development of logic-based convergent medical terminology. Methods Inf Med 1998;37:426–39.
[29] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007;25:1251–5.
[30] Payne PRO, Mendonça EA, Johnson SB, Starren JB. Conceptual knowledge acquisition in biomedicine: a methodological review. J Biomed Inform 2007;40:582–602.
[31] Bruce GB, David CW. Readings in knowledge acquisition and learning: automating the construction and improvement of expert systems. In: Bruce GB, David CW, editors. Morgan Kaufmann Publishers Inc.; 1993. p. 906.
[32] Buitelaar P, Cimiano P, Magnini B. Ontology learning from text: method, evaluation and applications. Amsterdam, Berlin, Oxford, Tokyo, Washington, DC: IOS Press; 2005.
[33] Navigli R, Velardi P, Gangemi A. Ontology learning and its application to automated terminology translation. IEEE Intell Syst 2003;18:22–31.
[34] Fensel D, Studer R. Knowledge acquisition, modeling and management. Springer; 2008.
[35] Shadbolt N, O'hara K, Schreiber G. Advances in knowledge acquisition. Springer; 2008.
[36] Hoffmann A. Advances in knowledge acquisition and management: Pacific Rim knowledge acquisition workshop. China: Springer Guilin; 2006.
[37] Hearst MA. Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 12th Conference on Computational Linguistics, 1992.
[38] Harris ZS. Mathematical structures of language. New York, NY, USA: Krieger Pub. Co.; 1968.
[39] Firth JR. Papers in linguistics. London: Oxford University Press; 1934–1957.
[40] Gamallo P, Agustini A, Lopes G. Selection restrictions acquisition from corpora. In: Proceedings EPIA. Springer; 2001.
[41] Grefenstette G. Explorations in automatic thesaurus discovery. Boston, MA: Kluwer Academic Publisher; 1994.
[42] Church KW, Hanks P. Word association norms, mutual information, and lexicography. In: Proceedings of 27th Annual Meeting of the ACL, 1989, p. 76–83.
[43] Smadja F. Retrieving collocations from text: xtract. Comput Linguist 1993;19:143–77.
[44] Caraballo S, Charniak E. Determining the specificity of nouns from text. In: Proceedings of SIGDAT, 1999.
[45] Lin D. Automatic retrieval and clustering of similar words. In: Proceedings of COLING, 1998.
[46] Hindle D. Noun classification from predicate-argument structures. In: Proceedings of 28th ACL, 1990, p. 268–275.
[47] Reinberger M-L, Spyns P. Unsupervised text mining for the learning of DOGMA-inspired ontologies. In: Proceedings of ECAI and EKAW, 2004.
[48] Agirre E, Ansa O, Hovy E, Martínez D. Enriching very large ontologies using the WWW. In: Proceedings of the Ontology Learning Workshop, Berlin, Germany, 2000.
[49] Gulla JA, Brasethvik T, Kvarv GS. Association rules and cosine similarities in ontology relationship learning, enterprise information systems. Berlin, Heidelberg: Springer; 2009. p. 201–212.
[50] Cherfi H, Toussaint Y. How far association rules and statistical indices help structure terminology? In: Proceedings of the15th ECAI: Workshop on Machine Learning and Natural Language Processing for Ontology Engineering. France: Lyon; 2002.
[51] Bodenreider O, Aubry M, Burgun A. Non-lexical approaches to identifying associative relations in the gene ontology. Pac Symp Biocomput 2005;2:91–102.
[52] Collier N, Nobata C, Tsujii J. Extracting the names of genes and gene products with a Hidden Markov Model. In: Proceedings of COLING, Sarrebruck, 2000.
[53] Morgan A, Hirschman L, Yeh A, Colosimo M. Gene name extraction using FlyBase resources. In: Proceedings of the ACL workshop on Natural language processing in biomedicine, 2003, p. 1–8.
[54] Shen D, Zhang J, Zhou G, Su J, Tan CL. Effective adaptation of Hidden Markov model-based named entity recognizer for biomedical domain. In: Proceedings of the ACL Workshop on Natural Language Processing in Biomedicine, 2003, p. 49–56.
[55] Kazamay JI, Makinoz T, Ohta Y, Tsujiiy JI. Tuning support vector machines for biomedical named entity recognition. In: Proceedings of the ACL workshop on Natural Language Processing in Biomedicine, 2003, p. 1–8.
[56] Yamamoto K, Kudo T, Konagaya A, Matsumoto Y. Protein name tagging for biomedical annotation in text. In: Proceedings of the ACL workshop on Natural Language Processing in Biomedicine, 2003, p. 65–72.
[57] Alfonseca E, Manandhar S. Extending a lexical ontology by a combination of distributional semantics signatures. In: Proceedings of EKAW, 2002, p. 1–7.
[58] Alfonseca E, Manandhar S. An unsupervised method for general named entity recognition and automated concept discovery. In: Proceedings of the 1st International Conference on General WordNet, 2002.
[59] Hasting PM. Automatic acquisition of word meaning from context. Comp. Sci. Eng, Univ. of Michigan; 1994.
[60] Hahn U, Schnattinger K. Towards text knowledge engineering. In: Proceedings of AAAI98, IAAI98 1998, p. 524–531.
[61] Basili R, Pazienza MT, Velardi P. An empirical symbolic approach to natural language processing. J Artificial Intell 1996;85:59–99.
[62] Hamon T, Nazarenko A. Detection of synonymy links between terms: experiment and results. In: Bourigault D, Jacquemin C, L'Homme M-C, editors. Recent Advances in Computational Terminology. John Benjamins Publishing Company; 2001, p. 185–208.
[63] Downey D, Etzioni O, Soderland S, Weld DS. Learning text patterns for Web information extraction and assessment. In: Proceedings of the American Association for Artificial Intelligence Workshop on Adaptive Text Extraction and Mining, 2004.
[64] Moldovan DI, Girju R. An Interactive tool for the rapid development of knowledge bases. International Journal on Artificial Intelligence Tools; 1999.

[65] Grefenstette G. Automatic thesaurus generation from raw text using knowledge-poor techniques, Ninth Annual Conference of the UW Centre for the New OED and text Research – Making Sense of Words, 1993.
[66] Geffet M, Dagan I. The distributional inclusion hypotheses and lexical entailment. In: Proceedings of the 43rd Annual Meeting of the ACL, 2005, p. 107–114.
[67] Faatz A, Steinmetz R. Ontology enrichment with texts from the WWW. Helsinki, Finland: Semantic Web Mining Workshop; 2002.
[68] Bikel D, Miller S, Schwartz L, Wesichedel R. Nymble: a high-performance learning name-finder. In: Proceedings of the Fifth Conference on Applied Natural Language Processing, 1997, p. 194–201.
[69] Chanlekha H, Collier N. A methodology to enhance spatial understanding of disease outbreak events reported in news articles. Int J Med Informatics 2010;79:284–96.
[70] Caraballo S. Automatic construction of a hypernym-labeled noun hierarchy from text. In: Proceedings of the 37th Conference on Computational Linguistics, 1999.
[71] Cederberg S, Widdows D. Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In: Proceedings of the 7th Conference on Natural Language Learning, 2003, p. 111–118.
[72] Fiszman M, Rindflesch TC, Kilicoglu H. Integrating a hypernymic proposition interpreter into a semantic processor for biomedical texts. In: Proceedings of the Annual Symp. of American Medical Informatics Association, 2003, p. 239–243.
[73] Snow R, Jurafsky D, Ng AY. Learning syntactic patterns for automatic hypernym discovery, Advances in Neural Information Processing Systems, 2004.
[74] Riloff E. Automatically generating extraction patterns from untagged text. In: Proceedings of the 13th National Conference on Artificial Intelligence, 1996.
[75] Velardi P, Navigli R, Cucchiarelli A, Neri F. Evaluation of OntoLearn, a methodology for automatic learning of domain ontologies. In: Proceedings of ECAI and EKAW, 2004.
[76] Cimiano P, Pivk A, Schmidt-Thieme L, Stabb S. Learning taxonomic relations from heterogenerous sources of evidence. In: Proceeding of EKAW, 2004.
[77] Rinaldi F, Yuste E, Schneider G, Hess M, Roussel D. Exploiting technical terminology for knowledge management. In: Proceedings of ECAI and EKAW, 2004.
[78] Morin E, Jacquemin C. Automatic acquisition and expansion of hypernym links. Comp Human 2004;38:343–62.
[79] Bodenreider O, Rindflesch TC, Burgun A. Unsupervised, corpus-based method for extending a biomedical terminology. In: Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain, 2002, p. 53–60.
[80] Ryu P-M, Choi K-S. Measuring the specificity of terms for automatic hierarchy construction. In: Proceedings of the ACL-SIGLX Workshop on Deep Lexical Acquisition, Ann Arbor, Michigan, June, 2005.
[81] Witschel HF. Using decision trees and text mining techniques for extending taxonomies. In: Proceedings of Learning and Extending Lexical Ontologies by using Machine Learning Methods, Workshop at ICML, 2005.
[82] M. Berland, E. Charniak, Finding parts in very large corpora. In: Proceedings of the 37th Conference on Computational Linguistics, 1999, pp. 57-64.
[83] Sundblad H. Automatic acquisition of hyponyms and meronyms from question corpora. In: Proceedings of the 15th European Conference on Artificial Intelligence. France: Lyon; 2002.
[84] Girju R, Badulescu A, Moldovan D. Learning semantic constraints for the automatic discovery of part-whole relations. In: Proceedings of the Human Language Technology Conference, 2003.
[85] Nenadć G, Spasić I, Ananiadou S. Automatic discovery of term similarities using pattern mining, COLING on COMPUTERM. In: 2nd International Workshop on Computational Terminology ACL, 2002, p. 1–7.
[86] Kavalec M, Svatek V. A study on automated relation labeling in ontology learning. In: Buitelaar P, Cimiano P, Magnini B, editors. Ontology learning from text: method, evaluation and applications. Amsterdam, Berlin, Oxford, Tokyo, Washington, DC: IOS Press; 2005. p. 44–58.
[87] Blaschke C, Valencia A. Automatic ontology construction from the literature. Genome Inform 2002;13:201–13.
[88] Turney PD. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In: Proceedings of the 12th European Conference on Machine Learning, 2001.
[89] Harris ZS. A grammar of English on mathematical principles. New York: Wiley; 1982.
[90] Harris ZS. A theory of language and information: a mathematical approach. Oxford: Clarendon Press; 1991.
[91] Friedman C, Alderson PO, Austin J, Cimino JJ, Johnson SB. General natural language text processor for clinical radiology. JAMIA 1994;1:161–74.
[92] Sager N, Lyman M, Buchnall C, Nhan NT, Tick LJ. Natural language processing and representation of clinical data. JAMIA 1994;1:142–60.
[93] GENIA:Available from: http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi.
[94] Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of ICML, 2001, p. 282–289.
[95] Riloff E, Shepherd J. A corpus-based approach for building semantic lexicons. In: Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP), 1997.
[96] Roark B, Charniak E. Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. In: Proceedings of ACL, 1998, p. 1110–1116.
[97] Widdows D, Dorow B. A graph model for unsupervised lexical acquisition. In: 19th International Conference on Computational Linguistics, 2002, p. 1093–1099.
[98] Deerwester S, Dumais S, Furnas G, Landauer T, Harshman R. Indexing by latent semantic analysis. J Am Soc Inform Sci 1990;41:391–407.
[99] Baeza-Yates R, Ribiero-Neto B. Modern information retrieval. Boston, MA: Addison Wesley/ACM Press; 1999.
[100] Rindflesch TC, Rajan J, Hunter L. Extracting molecular binding relationships from biomedical text. In: Proceedings of the 6th Applied Natural Language Processing Conference, ACL, 2000, p. 188–195.
[101] Rindflesch TC, Tanabe L, Weinstein JN, Hunter L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. In: Proceedings of PSB, 2000, p. 514–525.
[102] Riloff E. Automatically Constructing a Dictionary for Information Extraction Tasks. In: Proceedings of the Eleventh National Conference on Artificial Intelligence, AAAI Press, 1993, p. 811–816.
[103] Thelen M, Rilloff E. A Bootstrapping method for learning Semantic lexicons using Extraction Patterns Contexts. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002.
[104] Markó K, Schulz S, Hahn U. Automatic lexeme acquisition for a multilingual medical subword thesaurus. Int J Med Inform 2007;76:184–9.
[105] Cover TM, Thomas JA. Elements of information theory. New York, NY: John Wiley and Sons Inc.; 1991.
[106] Witschel H, Terminologie-extraktion – Möglichkeiten der Kombination statistischer und musterbasierter Verfahren. Würzburg: Ergon Verlag, 2004.
[107] Kodratoff M. Comparimg machine learning and knowledge discovery in databases: an application to knowledge discovery in text, ECCAI summer course, 1999.
[108] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases, Proceedings of the SIGMOD international conference on management of data, ACM, Washington, DC, United States, 1993, p. 207–216.
[109] Borgelt C. Efficient implementations of Apriori and Eclat, Proceedings of CEUR Workshop, Aachen, Germany, 2003.
[110] Grefenstette G. Sextant: exploring unexplored contexts for semantic extraction from syntactic analysis. In: Proceedings of the 30st annual meeting of the Association for Computational Linguistics, 1992.
[111] Wilbur WJ, Yang Y. An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. Comput Biol Med 1996;26:209–22.
[112] Faure D, Nedellec C, Rouveirol C. Acquisition of semantic knowledge using machine learning method: The System ASIUM. Technical Report #ICS-TR-88-16, Laboratoire de Recherche en Informatique, University Paris Sud (1998).
[113] Peat HJ, Willet P. The limitations of term co-occurrence data for query expansion in document retrieval systems. J Am Soc Inform Sci 1991;42:378–83.
[114] Diday E, Introduction a L'analyse des Donnees Symboliques, INRIA No. 1074, 1989.
[115] Faure D, Nedellec C. ASIUM: learning subcategorization frames and restrictions of selection. ECML Workshop on text mining, 1998.
[116] Yamaguchi T. Acquiring conceptual relationships from domain-specific texts. In: Proceedings of IJCAI Workshop on Ontology Learning (OL). USA: Seattle; 2001.
[117] Hearst MA, Schutze H, Customizing a lexicon to better suit a computational task. In: Proceedings of the ACL SIGLEX Workshop on Acquisition of Lexical Knowledge from Text, Columbus, OH, 1993.
[118] Shamsfard M, Barforoush AA. Learning ontologies from natural language texts. Int J Hum Comput Stud 2004;60:17–63.
[119] Etzioni O, Cafarella M, Downey D, Kok S, Popescu A-M, Shaked T, Soderland S, Weld DS, Yates A. Web Scale information extraction in know it all (preliminary results). In: Proceedings of the 13th International World Wide Web Conference, New York, USA, 2004, p. 100–111.
[120] Hahn U, Romacker M, Schulz S. Creating knowledge repositories from biomedical reports: the MEDSYNDIKATE text mining system. Pac Symp Biocomput 2002:338–49.
[121] Hahn U, Romacker M, Schulz S. Discourse structures in medical reports-watch out! The generation of referentially coherent and valid text knowledge bases in the MEDSYNDIKATE System. Int J Med Inform 1999;53:1–28.
[122] Strube M, Hahn U. Functional centering: grounding referential coherence in information structure. Comput Linguist 1999;25:309–44.
[123] Šarić J, Jensen LJ, Ouzounova R, Rojas I, Bork P. Extraction of regulatory gene/protein networks from Medline. Bioinformatics 2006;22:645–50.
[124] Maedche A, Volz R. The ontology extraction & maintenance framework Text-To-Onto. In: Proceedings of the ICDM Workshop on the Integration of Data Mining and knowledge management, San Jose, CA, USA, November 31, 2001.
[125] Srikant R, Agrawal R. Mining generalized association rules. Future Generation Computer Systems 1997;13:161–80.
[126] Kietz JU, Maedche A, Volz R. A method for semi-automatic ontology acquisition from a corporate intranet. In: Proceedings of Workshop Ontologies and Text, 2000.
[127] Maedche A, Staab S. Discovering conceptual relations from text. In: Proceedings of the 14th European Conference on Artificial Intelligence, Berlin, 2000, p. 321–325.

[128] Maedche A, Staab S. Semi-automatic engineering of ontologies from text. In: Proceeding of the 12th Internal Conference on Software and Knowledge Engineering. Chicago, USA, 2000.
[129] Maedche A, Staab S. Mining ontologies from text. In: Proceedings of EKAW, Springer Lecture Notes in Artificial Intelligence (LNAI-1937), 2000.
[130] Cimiano P, Völker J. Text2Onto – a framework for ontology learning and data-driven change discovery. In: Montoyo A, Muńoz R, Métais E, editors. NLDB. Heidelberg Alicante, Spain: Springer; 2005. p. 227–38.
[131] Mima H, Ananiadou S, Nenadić G, Tsujii J. A methodology for terminology-based knowledge acquisition and integration. Sarrebruck: Proceedings of COLING; 2000. pp. 667–673.
[132] Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms: the C-value/NC-value method. Int J Digit Libr 2000;3:115–30.
[133] Ushioda A. Hierarchical clustering of words. Sarrebruck: Proceedings of COLING; 1996.
[134] Craven M, DiPasquo D, Freitag D, McCallum A, Mitchell T, Nigam K, et al. Learning to construct knowledge bases from the world wide web. Artif Intell 2000;118:69–113.
[135] Ogren PV, Cohen KB, Acquaah-Mensah GK, Eberlein J, Hunter L. The compositional structure of Gene Ontology terms. Pac Symp Biocomput 2004:214–25.
[136] Ogren PV, Cohen KB, Hunter L. Implications of compositionality in the gene ontology for its curation and usage. Pac Symp Biocomput 2005:174–85.
[137] Spackman KA, Campbell KE. Compositional concept representation using SNOMED: towards further convergence of clinical terminologies. In: Proceedings of AMIA Annual Symp, 1998, p. 740–744.
[138] Pakhamov S, Coden A, Pakhomov S, Ando R, Duffy P, Chute C. Domain-specific language models and lexicons for tagging. J Biomed Inform 2005;38:422–30.
[139] Stetson PD, Johnson SB, Scotch M, Hripcsak G. The sublanguage of cross-coverage. In: Proceedings of AMIA Annual Symp., 2002, p. 742–746.
[140] Taira RK, Soderland SG, Jakobovits RM. Automatic structuring of radiology free-text reports. Radiographics 2001;21:237–45.
[141] Schadow G, McDonald CJ. Extracting structured information from free text pathology reports. In: Proceedings of AMIA Annual Symp., 2003, p. 584–588.
[142] Liu K, Chapman WW, Savova G, Chute CG, Sioutos N, Crowley RS. Effectiveness of lexico-syntactic pattern matching for ontology enrichment with clinical documents. Methods Inf Med, submitted for publication.
[143] Embarek M, Ferret O. Learning patterns for building resources about semantic relations in the medical domain. In: Proceedings of the 6th International Language Resources and Evaluation. Morocco: Marrakech; 2008.
[144] ODIE toolkit, 2010. Available from: http://bioontology.org/tools/ODIE.html.
[145] Guarino N, Welty CA. An overview of OntoClean. Handbook on Ontology 2004:151–72.
[146] Faatz A, Steinmetz R. An evaluation framework for ontology enrichment. In: Proceedings of ECAI and EKAW, 2004.