Formative Evaluation of Ontology Learning Methods for Entity Discovery by Using Existing Ontologies as Reference Standards

K. Liu[1]; K. J. Mitchell[1]; W. W. Chapman[2]; G. K. Savova[3]; N. Sioutos[4]; D. L. Rubin[5]; R. S. Crowley[1,6]

[1]Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA;
[2]Division of Biomedical Informatics, University of california San Diego, San Diego, CA, USA;
[3]Childrens´ Hospital Boston and Harvard Medical School, Boston, MA, USA;
[4]Lockheed Martin Corporation, Fairfax, VA, USA;
[5]Department of Radiology, Stanford University, Stanford, CA, USA;
[6]Department of Pathology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA

Keywords
Ontology learning from text, statistical ontology learning method, statistical ontology, learning algorithm, ontology enrichment, natural language processing, ontology evaluation

Summary
Objective: Developing a two-step method for formative evaluation of statistical Ontology Learning (OL) algorithms that leverages existing biomedical ontologies as reference standards.
Methods: In the first step optimum parameters are established. A 'gap list' of entities is generated by finding the set of entities present in a later version of the ontology that are not present in an earlier version of the ontology. A named entity recognition system is used to identify entities in a corpus of biomedical documents that are present in the 'gap list', generating a reference standard. The output of the algorithm (new entity candidates), produced by statistical methods, is subsequently compared against this reference standard. An OL method that performs perfectly will be able to learn all of the terms in this reference standard. Using evaluation metrics and precision-recall curves for different thresholds and parameters, we compute the optimum parameters for each method. In the second step, human judges with expertise in ontology development evaluate each candidate suggested by the algorithm configured with the optimum parameters previously established. These judgments are used to compute two performance metrics developed from our previous work: Entity Suggestion Rate (ESR) and Entity Acceptance Rate (EAR).
Results: Using this method, we evaluated two statistical OL methods for OL in two medical domains. For the pathology domain, we obtained 49% ESR, 28% EAR with the Lin method and 52% ESR, 39% EAR with the Church method. For the radiology domain, we obtain 87% ESA, 9% EAR using Lin method and 96% ESR, 16% EAR using Church method.
Conclusion: This method is sufficiently general and flexible enough to permit comparison of any OL method for a specific corpus and ontology of interest.

Correspondence to:
Kaihong Liu, MD, PhD
Center for Dental Informatics
Salk Hall, Suite 378
3501 Terrace Street
Pittsburgh, PA 15232
USA
E-mail: kaihong@pitt.edu

Methods Inf Med 2013; 52: 308–316
doi: 10.3414/ME12-01-0029
received: April 4, 2012
accepted: February 2, 2013
prepublished: May 13, 2013

1. Introduction and Background

Despite the increasing importance of domain ontologies in biomedical research, there remain significant barriers to their development. To be most useful, domain ontologies must achieve a high degree of coverage of the both entities and entity relationships. However, ontology development is typically a manual, time-consuming, and frequently error-prone process. Limited resources result in missing entities and relationships as well as difficulty in updating the ontology as knowledge changes. Ontology Learning (OL) from text uses methods developed in the fields of Natural Language Processing (NLP), Artificial Intelligence (AI), and Machine Learning (ML) in an attempt to minimize that human effort and improve coverage. Authors of this paper have previously published a literature review describing the research and advances of NLP methods and systems for ontology learning [1]. In general, these methods can be grouped into three different approaches: symbolic, statistical, and a hybrid of both. One widely known symbolic NLP approach is Hearst's pattern matching method [2]. Hearst used Lexical Syntactic Patterns (LSPs) to discover hyponyms from texts. For example, the LSP "such as" can be used to extract "benign eccrine neoplasia" as a hypernym for "nodular hidroadenoma" from the sentence: "Compatible with benign eccrine

neoplasia, such as nodular hidroadenoma." Berland and Charniak [3] used alternative patterns to find instances of "part-of" relationships. Statistical approaches have become more popular following a resurgence of interest in Harris's distributional hypothesis [4]. Researchers such as Church [5], Agirre [6], and Cimiano and Staab [7] have explored statistical approaches for new entity and relationship discovery in a variety of domains other than biomedicine, and have achieved promising results. Subsequently, many researchers in the biomedical domain have proposed to experiment with these approaches for biomedical ontology learning using free-text clinical documents.

Previous attempts to use ontology learning methods in biomedical knowledge acquisition have yielded mixed results. One important problem is that many methods that were developed specifically in other domains have not been tested and evaluated in the biomedical domain. Evaluating ontology learning methods is, in itself, a difficult task. Currently, there are four common approaches taken: *application-based evaluation; data-driven evaluation; domain-expert evaluation; and reference-standard, or gold standard-based evaluation. Application-based evaluation* measures the performance of an application based on the learned ontology [8–10], but only indirectly evaluates the learned ontology and is therefore not useful for algorithm development specifically focusing on learning the ontology itself. *Data-driven evaluation* [11] measures the fitness of an ontology to a domain by measuring the coverage of the domain corpus, but has disadvantages similar to those of application-based evaluation, and is not directly informative for OL algorithm development. *Domain-expert evaluation* [12] uses human judges to assess the accuracy of output of the OL method or system, and is considered to be a superior method – but it is typically time-intensive, and thus, expensive. *Reference-standard, or gold standard-based evaluation* [13–16] is the most direct method available if the reference standard has been established. It has the advantage of evaluating several levels of the learned ontology specifications (e.g. lexical, taxonomic, and relational). It also provides a method for comparing the different ontology-learning approaches and ontology-based tools, and is well suited for ontology-learning algorithm development. However, it often assumes that the reference standard adequately represents the domain knowledge and is complete, which, in many instances, is not the case.

Our two-step evaluation process combines the reference standard-based approach and the domain-expert approach for the formative evaluation of one or more ontology-learning algorithms, yielding the accuracy of domain-expert evaluation and the automation of reference-standard based evaluation. We have used a similar approach in our previous work [17], in which we described a two-step human method for evaluating purely symbolic/syntactic OL: the Hearst Lexical Syntactic Pattern (LSP) matching method, for biomedical domain ontology development. In the first step, domain judges identified the meaningful medical terms on either side of the lexical pattern in text documents. In the second step, ontology developers evaluated the accuracy of the entity candidate and relationships. However, for OL methods that are either statistical or hybrid, the two-step human evaluation approach is inefficient. Given the wide range of possible parameters for any of these methods, the effort required for human evaluation could be enormous. We have therefore devised a variation of our previous approach that begins with automated parameter selection using the target ontology as a reference standard. In the second step of the evaluation, we can determine performance of each OL method using human judgments by ontologists, based on metrics described previously [17], including Entity Suggestion Rate, Entity Acceptance Rate, Relationship Suggestion Rate, and Relationship Acceptance Rate.

The statistics-based OL method requires tuning and testing of parameters before it can be used in the final implementation for a particular domain. This tuning and testing step is often called the "method development stage". Because there are multiple outputs for each parameter and combination of parameters tested, domain-expert evaluation is prohibitive at this stage, but automation can seldom be achieved due to lack of reference standards. Our approach attempts to solve this problem through use of an existing ontology. We first identify all the entities (instances or classes) of an existing ontology that have been mentioned in our clinical documents-learning resource. We would assume if these entities are removed from the ontology, they would be the entities that an ontology enrichment method attempts to discover. Therefore, this set of entities could be used as our reference standard for evaluating ontology enrichment methods under varying conditions during the development stage, in order to identify the most promising methods. There are three advantages of this approach: 1) it solves the shortage of reference standards; 2) the formative aspects of evaluation for parameter selection of statistical methods can be automated once the reference standards are established; and 3) the OL methods developed using this approach will be more likely to learn new entities that reside within the scope of the targeted ontology. Domain-expert evaluation of the optimized set of candidates is performed as a second step. To test this methodology, we selected two statistical OL methods for entity extraction using two genre of clinical documents. The objective of this research is to establish an evaluation method that can be used for rapid development of OL methods in biomedicine.

2. Materials and Methods

2.1 Clinical Corpora

Two types of clinical documents, surgical pathology reports and radiology reports, served as ontology learning resources for our study. The corpus of surgical pathology reports included a total of 852,764 documents; the corpus of radiology reports included a total of 209,997 documents. Both corpora were obtained from clinical information systems at the University of Pittsburgh Medical Center (UPMC), which includes a total of 18 hospitals. Both corpora were de-indentified to meet the requirements of HIPAA "safe harbor" (18). Use of the clinical corpora was approved by the University of Pittsburgh's Institutional Review Board (IRB# PRO07070252).

2.2 Targeted Biomedical Knowledge Resources

We selected two biomedical knowledge resources in active development that could benefit from ontology enrichment using clinical text. The National Cancer Institute Thesaurus (NCIT) [19] is a description logic-based ontology sponsored by the National Cancer Institute. It includes more than 75,000 key biomedical concepts in over 20 categories, including Disease, Abnormal Cell, Molecular Abnormality, Organism, and Biological Process. RadLex [20], sponsored by the Radiology Society of North American (RSNA), is a lexicon for the uniform indexing and retrieval of radiology information resources. It includes over 11,000 concepts in 12 categories, including Imaging Observation, Procedure, Characteristic, and Treatment.

2.3 Named Entity Recognition Engine

From the many Named Entity Recognition (NER) system algorithms available [21–23], we chose IndexFinder [23] because of its computational efficiency. The IndexFinder algorithm starts with the first word of the text and iterates incrementally over each word progressively activating all phrases that contain that word. Once all the words of a phrase have been accounted for, the iteration stops for that phrase and it is added as a NamedEntity. The algorithm leverages four data structures: 1) an encoding of available vocabulary phrases indexed by phrase identifier (PID), 2) a sorting of phrase PIDs first by word count and then alphabetically, an inverse hash table of word to PID set, and 3) a PID length table specifying boundaries for PIDs of the same length The PID length table allows the algorithm to determine the number of words needed for a phrase detection without storing the number of words needed for all phrases. Using this technique, vocabularies as large as that of the Unified Medical Language System can be stored in memory on a PC with 500 megabytes of RAM.

We implemented the IndexFinder algorithm in Java as a UIMA processing resource for the ODIE system, and modified it by limiting the scope of the active phrase to sentences as opposed to the whole document. Like other commonly used mapping algorithms [21, 22], the method considers non-contiguous words. Given 1) a text corpus and 2) an OWL-formatted ontology, thesaurus, or lexicon as inputs, the IndexFinder NER system identifies named entities or their synonyms in a free-text corpus based on the knowledge resource(s) provided by the user (▶Figure 1 Part A). The knowledge resource (ontology, thesaurus, or taxonomy) can be changed by the user. From this system's output, we determine how many terms or synonyms in the corpus map to entities in the knowledge resource.

2.4 Establishing Reference Standards for Methodology Development

To establish reference standards for our study, A 'gap list' of entities is generated by finding the set of entities present in a later version of the ontology that are not present in an earlier version of the ontology. A named entity recognition system is used to identify entities in a corpus of biomedical documents that are present in the 'gap list', generating a reference standard (NCIT for the pathology corpus and RadLex for the radiology corpus). We assumed that if these entities did not already exist in the ontology, then we would discover them from the corpus using statistical methods, and the entire list would constitute the upper-bound of entities that we could discover. These lists became our reference standards for each domain. Once we had the reference standard, traditional evaluation metrics, such as recall and precision, could be employed.

2.5 Statistical Methods

To test this approach, we selected two statistical methods, Church's mutual information method [24] and Lin's similarity measure method [25], because they differed in their utilization of syntactic features for similarity measure and because each of the methods was well known and commonly used in its field.

The Church method measures the degree of similarity between two words, x and y, by measuring their co-occurring information or mutual information (I) in a corpus. The mutual information (I) between words x and y in a corpus is defined as follows:

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x) * P(y)}$$

where P(x) is the probability of x, P(y) is the probability of y, and P(x, y) is the joint probability of x and y. If there is a genuine association between x and y, then the joint probability P(x, y) will be much higher than chance P(x) · P(y), and consequently I(x, y) >> 0. If there is no relationship between x and y, then P(x, y) ≈ P(x) · P(y); thus, I(x, y) = 0. If x and y have complementary distribution, P(x, y) will be much less than P(x) · P(y) and I(x, y) << 0. P(x) and P(y) can be estimated using the frequency of the appearance of a word in a corpus. For x and y to be considered highly associated, there are two variables that can influence the results: one is the window size (WS) in which the two terms appear, and the other is the threshold (T) of the similarity of the two terms. The WR is the word distance between two terms, and the T is the cutoff point of the similarity score between two terms. If the similarity score is higher than the T, then the two terms are considered to be similar.

The Lin method uses syntactic information in addition to word co-occurrence information. The co-occurrence information between two words (w, w') and their grammatical relationship (r) are collected as the dependency triples (w, r, w'). For example, in the sentence "Patient had a high fever", the following set of dependency triples can be obtained: (had subject Patient); (had object fever); (fever adjective-modifier high); and (fever determiner a). In another sentence "The mucosa does not reveal any small polyps or other mass lesions", the dependency triples would be: (Reveal subject Mucosa), (reveal object polys), (reveal object lesions), (polys determiner any), (polys conjunction lesions), (polys modifier small), (lesion noun mass). If $||w, r, w'||$ denotes the frequency count of the dependency triple in the parsed corpus, the similarity between words w_1 and w_2 such as

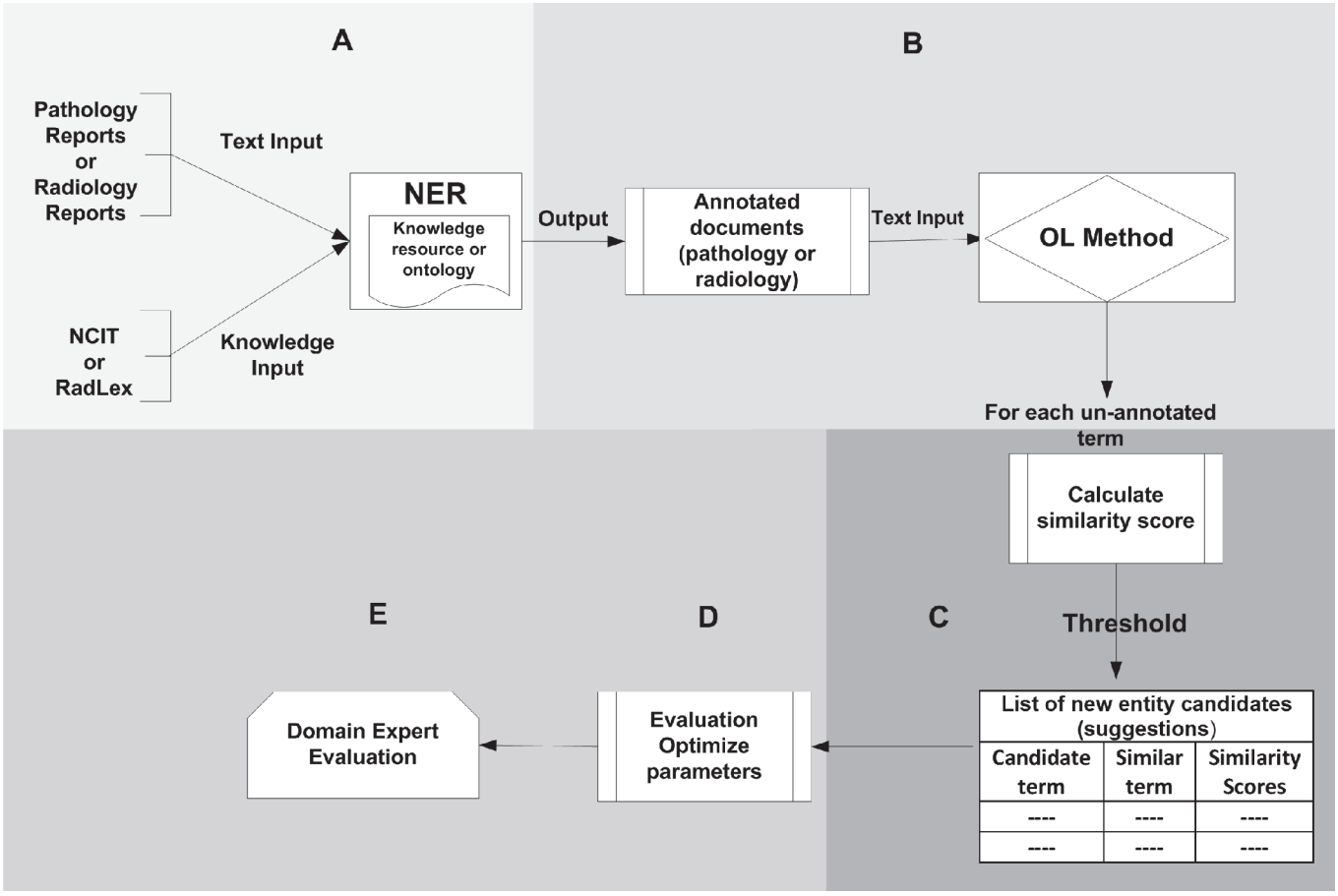

Figure 1 Work flow chart of new entity candidate extraction and evaluation using NER and statistical OL methods

"fever" and "polys" can be calculated as follows:

$$\text{Sim}(w_1, w_2) = \frac{\sum_{(r,w) \in T(w_1) \cap T(w_2)} (I(w_1, r, w) + I(w_2, r, w))}{\sum_{(r,w) \in T(w_1)} I(w_1, r, w) + \sum_{(r,w) \in T(w_2)} I(w_2, r, w)}$$

$$\text{where } I(w_1, r, w) = \log \frac{\| w_1, r, w \| * \| *, r, * \|}{\| w_1, r, * \| * \| *, r, w \|},$$

$$I(w_2, r, w) = \log \frac{\| w_2, r, w \| * \| *, r, * \|}{\| w_2, r, * \| * \| *, r, w \|}$$

The parameters for Lin's method include the two terms' appearance in differing syntactic relationships and the threshold of the similarity of the two terms. The method is representative of a hybrid of statistical and symbolic approaches to OL. Lin's MINIPAR [26] is used to parse the clinical corpus and generate dependency triples.

2.6 Generating New Entity Candidates

Both clinical document sets were divided into development and evaluation sets. We used the development set to experiment with the parameters for each OL method, while the evaluation set was set aside for final evaluation by domain experts. ▶Figure 1 illustrates the overall approach for generating new entity candidates (suggestions) and evaluation. First, the clinical documents were pre-processed by the Clinical Text Analysis and Knowledge Extraction System (cTAKES), an NLP engine [27]. However, the cTAKES dictionary lookup engine was replaced by our own NER engine, based on the IndexFinder algorithm, which maps entities in the corpus to the ontology. For each noun phrase that was not identified as an entity mention, the OL algorithm used the similarity measure described earlier to determine how similar it was to an existing entity in the ontology. The output of the methods consisted of lists of paired terms that were considered similar based on determinations from the statistical methods where one of the two terms was already in the ontology and the other was not. Terms were added to the list of new entity suggestions when similarity scores exceeded the threshold. Different thresholds produced different sets of entity suggestions.

We first used the NER system (▶Figure 1 Part A) to annotate the clinical corpora (pathology or radiology) with the ontologies NCIT v0909c or RadLex v3.03. The output of NER system consisted of sets of documents being annotated with entities that in turn became the input of OL algorithms (▶Figure 1 Part B). For each of the terms that has been annotated as an entity mention in the documents, the algorithm returned a list of similar terms, with similarity scores given in a descending order

Table 1 Syntactic relationships in each set explored with the Lin Method (obj: objective; mod: modifier; nn: noun; conj: conjunction; det: determiner; subj: subject; lex-mod: lexical modifier; pcomp-n: prepositional complement of the noun ; appo: appositional modifier; pred: predicate; punc: punctuation;)

	Syntactic Relationship Set										
	1	2		3	4	5	6	7	8	9	10	11
Syntactic Relationship (SR)	All SR	Top 10 most frequent SRs for NCIT	Top 10 most frequent SRs for RadLex	obj	mod	nn	conj	obj, mod, conj, nn	obj, mod, conj, nn, det	subj	obj, mod, conj, nn, det, subj	Lex-mod
mod	x	x	x		x			x	x		x	
nn	x	x	x			x		x	x		x	
pcomp-n	x	x	x									
subj	x	x	x							x	x	
conj	x	x	x	x			x	x	x		x	
obj	x	x	x					x	x		x	
punc	x	x	x									
lex-mod	x	x	x									x
det	x	x	x						x		x	
pred	x	x										
appo	x		x									

(▶Figure 1 Part C). The OL algorithm output is represented by a table in which the first column reveals the terms that have entity mappings of the ontology, the second column reveals the similar terms, and the third details the similarity scores. Terms would be added to the list of new entity candidates (suggestions) when similarity scores exceed a threshold that the user defines.

2.7 Evaluation Study Using Reference Standards

The traditional evaluation metrics, precision and recall, were used to measure the performance of the algorithms (▶Figure 1 Part D). If the term in the candidate list was also present in the reference standard, we considered it a true positive; otherwise, it was considered to be a false positive. For each set of suggestions, we calculated recall, precision, and F-score. Recall was computed as the total number of positive terms divided by the total number of terms in the reference standard. Precision was computed as the total number of true positives divided by the total number of candidates. F-score is the harmonic combination of recall and precision. Changing the threshold would alter the entity suggestion list. A precision-recall curve was drawn with the x-axis as 1– precision and the y-axis as recall for different thresholds. The best threshold was obtained based on the operating point on the curve which corresponds to the best F-score.

2.8 Generating Optimal Parameters

The application of statistical methods to OL requires parameter selection. The value of the parameters can vary the results dramatically, especially within different domains. Furthermore, the relative performance of different methods is best compared when each method is optimized, and when the methods can be applied under identical conditions (e.g., with the same corpus and knowledge resource). Parameters in OL methods encompass a wide range of variables. For example, in the Church method, the key parameter is window size. In this study, we experimented with 7 different window sizes (2, 3, 4, 5, 6, 7, and 15). The best combination of window size and threshold was obtained after we compared all the precision-recall curves. In contrast, the key parameter for the Lin method is the type of syntactic relationship. In this study, we experimented with 11 different sets of syntactic relationships (▶Table 1), and generated a precision-recall curve for each set. The best combination of syntactic relationship and threshold was obtained after we compared all precision-recall curves.

2.9 Domain-expert Evaluations

After we obtained suitable parameters for each method and domain, we proceeded to the final stage: evaluation by domain experts (▶Figure 1 Part E). A set of 30,000 clinical reports for each domain (pathology and radiology) was used as the learning resource. OL methods were run separately on each set of clinical documents using the optimal parameters we selected early in the method development stage. We randomly selected a subset of 100 pairs of terms from each output, and gave these data to the domain experts for final evaluation.

We invited two experienced curators to perform the final evaluations. One ontology curator, a pathologist currently working on the National Cancer Institute Thesaurus, evaluated the term list obtained from the surgical pathology corpus. The other, a radiologist who is currently curating RadLex, evaluated the term list obtained from the radiology corpus. For each

term in a term-pair, curators were asked the following questions:

1) Is the term already represented in the resource (possibly as a synonym)?
2) If not, should a new entity based on this term be added to the resource?
3) If not, what is the reason for the determination that it should not be added?

For each pair of terms, ontology curators also explored the following questions:

4) If there is a relationship between the two terms, what is the relationship? (Relationship choices were restricted to synonym, hypernym/hyponym, meronym, and other)
5) Does this relationship exist in the resource?
6) If not, should the relationship be added to the resource?
7) If no new relationship should be added, what is the reason for this determination?

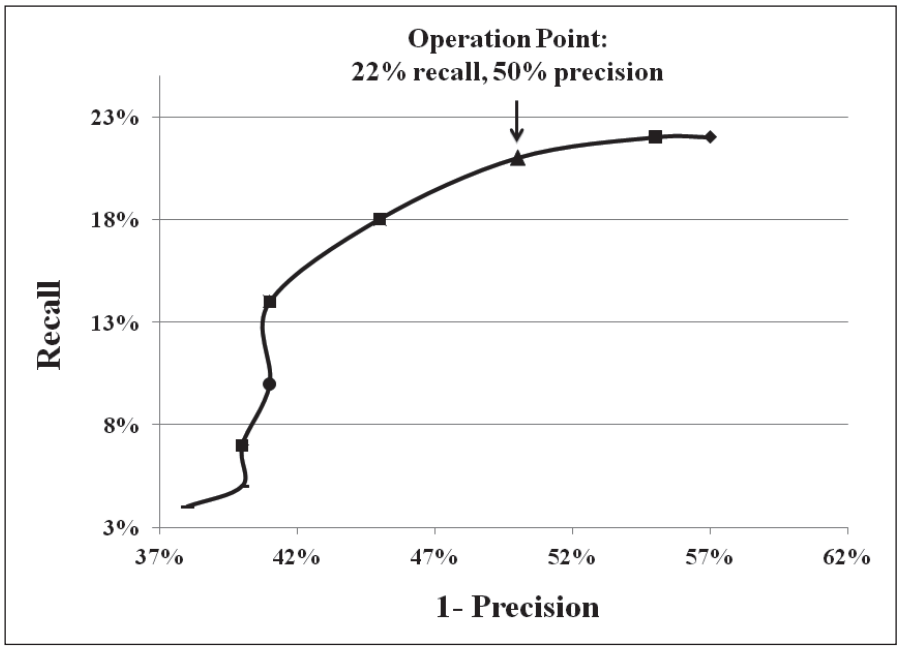

Figure 2 Precision-recall curve of Lin OL method for NCIT enrichment using pathology reports for all syntactic relationships (set 1)

2.10 Defining Final Evaluation Metrics

The classic measure of precision is not entirely adequate in summarizing the resulting data, since it does not capture the two-step process we anticipate using for suggesting new ontological elements. Therefore, we previously defined more specific evaluation metrics to quantify efficacy for the two discrete steps [17], but now update the definitions to encompass entities as opposed to concepts.

Entity Suggestion Rate (ESR):

$$ESR = \frac{\#\text{ of terms that were not in the ontology}}{\text{Total \# of terms extracted by the method}}$$

This metric indicates the percentage of terms, extracted using the enrichment method, that are new entity candidates and would be presented to the curator for a given target ontology.

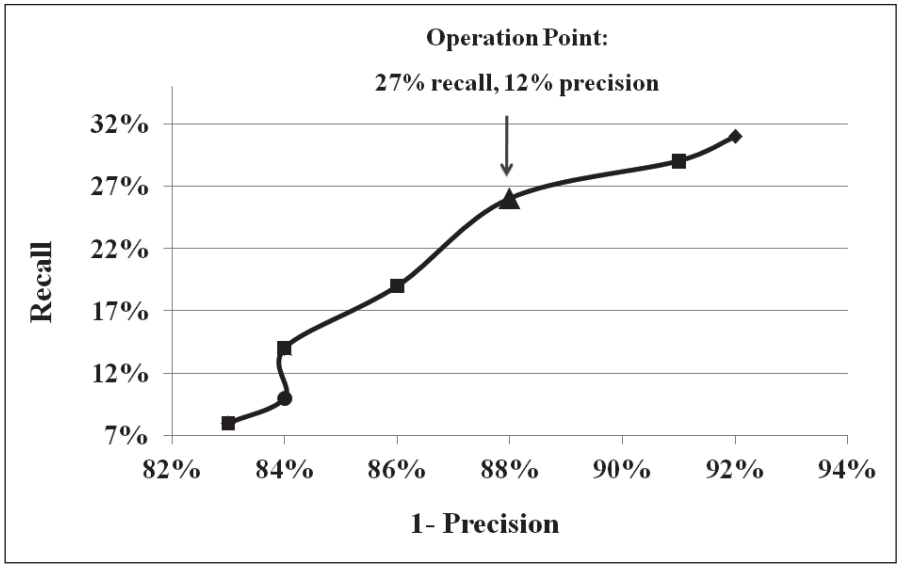

Figure 3 Precision-recall curve of Lin OL method for RadLex enrichment using radiology reports for all syntactic relationships.

Entity Acceptance Rate (EAR)(▶ Figure 4)

$$EAR = \frac{\#\text{ of terms that should be included as new concept, instance, or synonym in the ontology}}{\text{Total \# of terms extracted by the method that were not in the ontology}}$$

Figure 4 Entity Acceptance Rate (EAR)

Syntactic Relationship Set	NCIT				RadLex			
	T	Recall	Precision	F	T	Recall	Precision	F
1 (all relationships)	0.07	0.22	0.48	0.15	0.09	0.27	0.12	0.08
2 (top ten frequent relationships)	**5.46**	**0.22**	**0.51**	**0.15**	–2.63	0.05	0.03	0.04
7 (obj, mod, conj, nn)	*0*	*0.22*	*0.42*	*0.15*	3.4	0.28	0.11	0.08
8 (obj, mod, conj, nn, det)	0.1	0.22	0.43	0.15	**7.00**	**0.25**	**0.13**	**0.09**

Table 2 Evaluation results for Lin OL method comparing different domains (T: threshold; F: F- score).

Table 3 Evaluation results for Church OL method comparing different domains

Window Size	NCIT				RadLex			
	T	Recall	Precision	F	T	Recall	Precision	F
2	15.74	0.36	0.54	0.22	16.38	0.61	0.04	0.04
3		0.46	0.48	0.23		0.64	0.04	0.04
4		0.42	0.49	0.23		0.53	0.04	0.04
5		0.46	0.48	0.23		0.64	0.04	0.04
15		0.52	0.41	0.23		0.52	0.04	0.04

This metric indicates the percentage of terms, extracted using the enrichment method, that would be added to the relevant ontology; these may represent new entities or new instances.

3. Results

The entities in the 'gap list' that were also in the corpus included a total of 5,281 NCIT entities and a total of 660 RadLex entities. These served as our reference standards for ontology learning algorithms for each respective domain.

▶Figure 2 shows the precision-recall curve for all syntactic relationships developed by the Lin method (Set 1) for NCIT enrichment using pathology reports, while ▶Figure 3 shows the precision-recall curve for the Lin method using all syntactic relationships (set 1) for RadLex enrichment using radiology reports. In general, with increased threshold, recall increases and precision decreases.

For each values of the parameter (e.g. syntactic sets), we selected the best threshold based on the operating point on the curve which is the best F-score (the harmonic combination of recall and precision). As shown in ▶Figure 2, the arrow designates the operating point that has 22% recall and 50% precision. The threshold for that point is 0.07. The operating point for ▶Figure 3 is at threshold 0.09, with 27% recall and 12% precision.

▶Table 2 shows the different combinations of syntactic relationships with operating points for enrichment of both NCIT and RadLex. It is evident that the top ten most frequent relationships (set 2) is the best combination of syntactic relationships for the Lin OL method using NCIT and pathology reports. With threshold 5.46, it achieved 22% recall and 51% precision. In contrast, the combination of object, modifier, conjunctive, noun phrase, and determiner (set 8) is the best combination of syntactic relationships for the Lin OL method using RadLex and radiology reports. It achieved 25% recall and 13% precision at a threshold of 7.0.

The same approach described above was used for investigating window size using the Church method; however, only final results are detailed below. ▶Table 3 shows the different window sizes with operating point for enrichment of both NCIT and RadLex. From this table, it is evident that window sizes 3 and 5 generate the best evaluation results, and are nearly identical. In this case, we selected window size 3 over window size 5, because the former was more computationally efficient.

▶Table 4 provides a summary of the best parameters selected for each method and domain, along with the recall and precision scores obtained. For the Church OL method, window size 3 and threshold 14.0 for NCIT are optimal, as are window size 3 and threshold 15.9 for RadLex. For the Lin OL method, syntactic relationship set 8 and threshold 7 produce the best results for NCIT; syntactic relationship set 2 and threshold 5.5 are preferable for RadLex. Overall, the Church method produced greater recall than the Lin (46% vs 22% for NCIT and 64% vs 25% for RadLex).

Table 4 Summary of evaluation results for OL methods in both domains

Domain	Church OL Method					Lin OL Method				
	Parameter		Evaluation Metrics			Parameter		Evaluation Metrics		
	WS	T	Recall	Precision	F	SR	T	Recall	Precision	F
NCIT	3	14.0	46%	48%	0.23	8 (obj, mod, conj, nn, det)	7	22%	51%	0.15
RadLex	3	15.9	64%	4%	0.04	2 (top 10 most frequent relationships)	5.5	25%	13%	0.09

3.1 New Entity Suggestion Rate and Acceptance Rate

Using the thresholds obtained during the development of the OL methods (▶Figure 1C), we were able to extract a total of 2,249 suggestions for NCIT and a total of 1,300 for RadLex using Lin's method. Among these, 49% of the 2,249 suggestions had not appeared in NCIT, and 87% of the 1,300 suggestions were absent from RadLex. Using Church's method, we attained a total of 4,529 suggestions for NCIT and 12,174 suggestions for RadLex. 52% of the 4,529 suggestions had not been found in NCIT and 96% of the 12,174 suggestions had not been found in RadLex. These results are represented in ▶Table 5 as ESRs. EARs were based on the domain experts' evaluations of a random sample of 100 terms culled from the suggestions. For NCIT, among the 100 sampled terms taken from the Lin method's extractions, 38 had not been included previously in the ontology; the domain experts determined that 28 of these 38 terms should be added, for a EAR of 28%. Among the 100 sampled terms taken from the Church extraction, 73 had not been included in the ontology; domain experts determined that 39 of the 73 terms should be added, for a EAR of 39% (▶Table 5). For RadLex, among the 100 terms sampled from the Lin extractions, 38 had not been found in the ontology; the domain experts determined that 9 of the 38 terms should be added, for a EAR of 9% (▶Table 5). We also manually examined the two lists of suggested terms for NCIT and two lists of suggested terms for RadLex by each method, and found that there were no overlapping terms.

Table 5 Entity suggestion and acceptance rates for the Lin and Church methods

	Pathology reports (Enrich NCIT)		Radiology reports (Enrich RADLex)	
OL Method	**ESR**	**EAR**	**ESR**	**EAR**
Lin Method	49% (1,100/2,249)	28% (28/100)	87% (1,135/1,300)	9% (9/100)
Church Method	52% (2,159/4,529)	39% (39/100)	96% (11,743/12,174)	16% (16/100)

4. Discussion

Despite significant developments in Ontology Learning methodology, such methods are only rarely applied to biomedical ontologies [28–30]. One of the major barriers to progress is the difficulty of evaluating the methods. The lack of systematic evaluation methods or reference standards makes it extremely difficult to compare algorithm performance among domains or for different knowledge resources. Further, the absence of reference standards for this task makes it particularly difficult to select parameters for statistical methods. We sought to address these problems by utilizing existing ontologies. For this study, we selected NCIT and RadLex as our ontology resources, to derive a set of reference standards that in turn allowed us to test, tune, and directly compare two statistical OL methods on clinical documents for recall, precision, and F-score. We found that the Church mutual information method performed better than Lin's similarity metric for both domains, with results of 46% vs. 22% recall for NCIT, and 64% vs. 25% recall for RadLex. The ESRs were 52% vs. 49% for NCIT and 96% vs 87% for RadLex and EARs were 39% vs 28% for NCIT and 16% vs. 9% for RadLex. The results appear to contradict common wisdom that addition of syntactic information should improve performance. However, it may be the case that syntactic rules work better for some cases or domain than others. Alternatively, this finding may relate to parsing inaccuracy due to differences between clinical documents and general English documents. To obtain the dependency triples, the entire corpus must be tagged for parts of speech (POS) and then parsed to generate the dependency triples. Differences between medical sublanguages and general English may produce inaccuracies in POS tagging and parsing that could affect downstream results. The finding confirms the importance of testing OL methods in the context of a targeted corpus and ontology.

In previous work, we described a two-step human method for evaluating purely symbolic/syntactic OL for biomedical domains. In the first step, domain judges identified the meaningful medical terms on either side of the lexical pattern in text documents. In the second step, ontology developers evaluated the accuracy of the entity candidate and relationships. For statistical or hybrid OL methods, the process is more complex. Because there are a wide range of parameters an OL method has to select when used for a certain domain, the effort required for human evaluation could be enormous. Therefore, we have devised a variation of our previous approach. We still used the two-step approach, however, the first step is the method development and aimed at automated parameter selection using the target ontology as a reference standard. One of the major issues of OL method development is the lack of reference standard. Our approach not only allows us to establish a reference standard in a rather quick manner but also allows the automation of formative aspect of the evaluation. There are added benefits as well: it provides for a ready comparison of OL method performance within a single domain and ensures that methods developed using this approach will be more likely to learn new entities that reside within the scope of the targeted ontology. Another potential benefit of using a reference standard for OL method evaluation is that it provides a systematic method of evaluation for multiple OL tasks, including learning of entities and taxonomic relationships.

We believe this methodology is sufficiently general and flexible enough to permit comparison of any OL method for a specific corpus and ontology of interest. However, we also see some limitations to this approach. First, a basic assumption is that the ontology or knowledge resource used to generate the reference standard is adequate and correctly represents the domain knowledge. Sometimes, this may not be the case. A second assumption is that the corpus used for ontology enrichment has some overlap with the target ontology. When this is not the case, the size of the reference standard may not be adequate for

OL method evaluation. We have found that we were only able to identify a total of 660 RadLex entities in the radiology corpus. This is perhaps the most important factor for poor precision results for RadLex enrichment (4% for Church's method and 13% for Lin's method).

5. Conclusion

This methodology provides a method for researchers interested in evaluating and using statistical and hybrid OL methods for the enrichment of biomedical ontologies or other knowledge resources. The approach can be used as the first of a two-step process in which the second step calls for human judgment by the ontology developer. All implementations of the methods (Church, Lin and IndexFinder) and source code used in this study are available as open source code on the Stanford NCBO gforge site for the Ontology Development and Information Extraction (ODIE) project [31].

Acknowledgments

The authors wish to thank Lucy Cafeo and Karma Lisa Edwards of the University of Pittsburgh for editorial assistance. This work was supported by NIH grant RO1 CA 127979.

References

1. Liu K, Hogan WR, Crowley RS. Natural Language Processing methods and systems for biomedical ontology learning. Journal of Biomedical Informatics 2011; 44 (1): 163–179.
2. Hearst MA. Automatic acquisition of hyponyms from large text corpora. Proceedings of the 12th Conference on Computational Linguistics; 1992; Nantes, France. 1992. pp 539–545.
3. Berland M, Charniak E. Finding parts in very large corpora. Proceedings of the 37th Conference on Computational Linguistics; 1999; College Park, MD. 1999. pp 57–64.
4. Harris ZS. Mathematical structures of language. New York, NY, USA: Krieger Pub Co; 1968.
5. Church KW, Hanks P. Word association norms, mutual information, and lexicography. Proceedings of 27th Annual Meeting of the Association for Computational Linguistics; 1989; Vancouver, BC, Canada. 1989. pp 76–83.
6. Agirre E, Ansa O, Hovy E, Martínez D. Enriching very large ontologies using the WWW. Proceedings of the Ontology Learning Workshop of the European Conference of AI (ECAI); 2000; Berlin, Germany. 2000.
7. Cimiano P, Pivk A, Schmidt-Thieme L, Stabb S. Learning taxonomic relations from heterogeneous sources of evidence. Proceeding of ECAI2004 Workshop on Ontology Learning and Evaluation, A Workshop at the 16th European Conference on Artificial Intelligence; 2004; Valencia, Spain. 2004.
8. Gliozzo A, Gioliano C, Strapparava C. Domain kernels for word sense disambiguation. Proceedings of the 43rd Annual Meeting of the ACL; 2005. 2005. pp 403–410.
9. Sánchez D, Moreno A. Web-scale taxonomy learning. Proceedings of the Workshop on Learning and Extending Lexical Ontologies by using Machine Learning; 2005. 2005.
10. Hartrumpf S. Extending knowledge and deepening linguistic processing for question answering. In: Peters C (ed). Results of the CLEF 2005 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2005 Workshop 2005. pp 361–369.
11. Brewster C, Alani H, Dasmahapatra S, Wilks Y. Data Driven Ontology Evaluation. International Conference on Language Resources and Evaluation (LREC 2004); Lisbon, Portugal. May 2004. pp 24–30.
12. Navigli R, Velardi P, Cucchiarelli A, Neri F. Automatic Ontology Learning: Supporting a Per-Concept Evaluation by Domain Experts. Workshop on Ontology Learning and Population (ECAI 2004), Valencia, Spain. 2004.
13. Brank J, Mladeniæ D, Grobelnik M. Gold standard based ontology evaluation using intance assigment. Proceedings of the 4th Workshop on Evaluating Ontologies for the Web (EON2006); 2006; Edinboro, Scotland.; 2006.
14. Dellschaft K, Staab S. On how to perform a gold standard based evaluation of ontology learning. In: I. F. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, et al., editors. International Semantic Web Conference, Lecture Notes in Computer Science 2006. Springer, 2006. pp 228–241.
15. Meadche A, Staab S. Measuring similarity between ontology. Proceedings of the European Conference on Knowledge Acquisition and Management; 2002. pp 251–263.
16. Maynard D, Peters W, Li Y. Metrics for evaluation of ontology-based information extraction. Proceedings of the EON 2006 Workshop; 2006; Edinburgh, UK; 2006.
17. 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 2010; 49 (6): 397–407.
18. Martins AL. Using grammar Iinference techniques in ontology learning. (Thesis) 2006.
19. National Cancer Institute Thesaurus (NCIT) 2010 (updated 2010; cited). Available from: http://ncit.nci.nih.gov/.
20. Mejino JLV, Rubin DL, Brinkley JF. FMA-RadLex: an application ontology of radiological anatomy derived from the Foundational Model of Anatomy reference ontology. Proceedings of the Annual Symposium of American Medical Informatics Association; Washington, DC. 2008. p 465.
21. MetaMap Technology Transfer (MMTX). 2012 (updated 2012; cited). Available from: http://ii.nlm.nih.gov/MMTx.shtml.
22. Dai M, Shah N, Xuan W, Musen M, Watson S, Athey B, et al. An Efficient Solution for Mapping Free Text to Ontology Terms. Mgrep. AMIA Summit on Translational Bioinformatics; San Francisco, CA 2008.
23. Zou Q, Chu W. IndexFinder: A Knowledge-based Method for Indexing Clinical Texts (cited). Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.841.
24. Church KW, Hanks P, editors. Word association norms, mutual information, and lexicography. Proceedings of 27th Annual Meeting of the ACL; 1989.
25. Lin D. Automatic retrieval and clustering of similar words. COLING '98 Proceedings of the 17th international conference on Computational linguistics; 1998; Montreal, Quebec, Canada.; 1998. pp 768–774.
26. Lin D. MINIPAR. 2012 (updated 2012; cited). Available from: http://webdocs.cs.ualberta.ca/~lindek/minipar.htm.
27. Clinical Text Analysis and Knowledge Extraction System (cTakes). 2012 (updated 2012; cited). Available from: http://ohnlp.sourceforge.net/cTAKES/.
28. Bodenreider O, Rindflesch TC, Burgun A. Unsupervised, corpus-based method for extending a biomedical terminology. Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain; 2002. pp 53–60.
29. Kazamay Ji, Makinoz T, Ohta Y, Tsujiiy Ji. Tuning support vector machines for biomedical named entity recognition. Proceedings of the ACL workshop on Natural language processing in biomedicine; 2003. pp 1–8.
30. Blaschke C, Valencia A. Automatic ontology construction from the literature. Genome Inform 2002. pp 201–213.
31. Ontology Development and Information Extraction (ODIE) 2012 (updated 2012; cited). Available from: https://bmir-gforge.stanford.edu/gf/project/odie.