Published in final edited form as:
Methods Inf Med. 2011 October 17; 50(5): 397–407. doi:10.3414/ME10-01-0020.

Effectiveness of Lexico-Syntactic Pattern Matching for Ontology Enrichment with Clinical Documents

K. Liu, MD, MS[1], **W.W. Chapman, PhD**[1,2], **G. Savova, PhD**[3], **C.G. Chute, MD, Dr.PH**[3], **N. Sioutos, MD**[4], and **R.S. Crowley, MD, MS**[1,2,5]

[1]Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, Pittsburgh, PA

[2]Intelligent Systems Program, University of Pittsburgh, Pittsburgh PA

[3]Department of Health Services Research, Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN

[4]Lockheed Martin Corporation, Fairfax, Virginia

[5]Department of Pathology, School of Medicine, University of Pittsburgh, Pittsburgh, PA

Summary

Objective—To evaluate the effectiveness of a Lexico-Syntactic Pattern (LSP) matching method for ontology enrichment using clinical documents.

Methods—Two domains were separately studied using the same methodology. We used radiology documents to enrich RadLex and pathology documents to enrich National Cancer Institute Thesaurus (NCIT). Several known LSPs were used for semantic knowledge extraction. We first retrieved all sentences that contained LSPs across two large clinical repositories, and examined the frequency of the LSPs. From this set, we randomly sampled LSP instances which were examined by human judges. We used a two-step method to determine the utility of these patterns for enrichment. In the first step, domain experts annotated Medically Meaningful Terms (MMTs) from each sentence within the LSP. In the second step, RadLex and NCIT curators evaluated how many of these MMTs could be added to the resource. To quantify the utility of this LSP method, we defined two evaluation metrics: Suggestion Rate (SR) and Acceptance Rate (AR). We used these measures to estimate the yield of concepts and relationships, for each of the two domains.

Results—For NCIT, the concept SR was 24%, and the relationship SR was 65%. The concept AR was 21%, and the relationship AR was 14%. For RadLex, the concept SR was 37%, and the relationship SR was 55%. The concept AR was 11%, and the relationship AR was 44%.

Conclusion—The LSP matching method is an effective method for concept and concept relationship discovery in biomedical domains.

Keywords

Ontology Learning from Text; Knowledge Acquisition; Ontology Enrichment; Natural Language Processing; Lexico-Syntactic Pattern

Correspondence and reprint requests to: Kaihong Liu, MD, MS, Department of Biomedical Informatics, UPMC Cancer Pavilion, Suite 301, 5150 Centre Avenue, Pittsburgh, PA 15232, kaihong@pitt.edu, telephone: (412) 647-6510, fax: (412) 647-5380.

INTRODUCTION

The development of biomedical ontologies represents a key advance of biomedical informatics during the past two decades [1–3]. Biomedical ontologies provide the foundation for system interoperability such as HL7 on the Reference Information Model [4]; are important elements of decision support systems [5–7], support clinical information retrieval [8,9]; and, are needed for natural language processing (NLP) tasks such as information extraction [10], anaphora resolution [11,12], and question answering [13]. Despite the critical role of ontologies in biomedicine, there remain many barriers to their widespread use. One well-known problem, termed the "knowledge acquisition bottleneck," is the extraordinary manual effort that is required to create and maintain these resources [14–16]. The fields of knowledge acquisition, ontology learning, and ontology learning from text provide methods for automated and semi-automated ontology enrichment, which may help reduce the burden of populating ontologies.

Knowledge acquisition is a broad field that encompasses the tasks of acquiring and structuring knowledge from a wide range of resources, including experts. Semi-automated and fully automated methods for knowledge acquisition use data that can be derived from structured data sources (e.g. databases), semi-structured sources (e.g. web pages) or completely unstructured sources (e.g. free text). Knowledge acquisition methods can be used to populate many kinds of knowledge representations. Ontology learning represents a subfield of knowledge acquisition that is specifically interested in extraction of ontological concepts and relationships from knowledge-rich resources. Ontology learning from text defines a more specific task that focuses exclusively on extraction of ontological elements from unstructured sources.

There are two major advantages of ontology learning from text in biomedical domain. First, the biomedical literature is an important mechanism for reporting new discoveries in biomedical science. MEDLINE, the largest and most widely used biomedical literature repository, contains approximately 16 million journal articles, and 2,000 to 4,000 new articles are added each day [17]. For several decades, researchers have used MEDLINE for the purpose of knowledge extraction [18–20]. Clinical documents provide a second document resource, representing more medically related domain knowledge, and they have become increasingly available in electronic formats. These documents have been utilized for many knowledge-based systems such as bio-surveillance systems [21,22]. Second, and more importantly, there is a direct connection between text and ontology because terms found in texts are linguistic representations or labels for concepts and relationships in an ontology [23–25]. There is increasing interest in the use of ontology learning from text to populate, maintain and update an ontology [26,27]. New concepts are often documented first within text, and the direct connection between text and ontology has made literature and documents preferable when choosing learning resources.

The biomedical research community has a long history of actively seeking and utilizing multiple methods for automatic extraction of semantic knowledge from free text within the disciplines of Natural Language Processing (NLP), Artificial Intelligence (AI) and Computational Linguistics (CL). These methods include: 1) linguistic or symbolic methods [28–31]; 2) corpus or statistical methods [32–34]; and, 3) hybrid methods [35–37]. In previous work, we reviewed methods and systems that are applicable to ontology learning in Biomedicine [38]. The long-term goal of our work is to investigate all of these classes of methods for use in our newly developed semi-automated ontology enrichment platform, Ontology Development and Information Extraction (ODIE) [39,40]. As a first step towards this long-term goal, we explored the use of Lexico-Syntactic Patterns (LSPs) for extracting related concepts and relationships between concepts from text. LSPs are surface relational

markers that exist in natural language. For example, in the phrase “systemic granulomatous diseases such as Crohn's disease or sarcoidosis,” the LSP “such as” can help us infer that “systemic granulomatous diseases” is a hypernym of “Crohn's disease” and “sarcoidosis”. The objective of the present study is to evaluate the effectiveness of using the LSP matching method for medical domain ontology enrichment using clinical reports.

I. BACKGROUND

Hearst [31] was the first to explore the LSP matching method for ontology learning of conceptual relationships. Hearst hypothesized that syntactic regularities within a specialized corpus reflect domain knowledge. For example, from the sentence “….works by authors ***such as*** Herrick, Goldsmith, and Shakespeare.”, the LSP “NP1 ***such as*** NP” suggests a hyponymic relationship between the noun phrase “authors” and the noun phrases “Herrick”, “Goldsmith”, and “Shakespeare”. Hearst searched for a set of predefined LSPs that indicated some relationship such as hyponym/hypernym in the text of the Grolier’s American Academic Encyclopedia. In this 8.6 million word encyclopedia, she found 7,067 sentences that contained the pattern ‘such as’. Out of these, 330 relationships were found. The advantages of this method as summarized by Hearst, are: 1) it does not require an extensive knowledge base; 2) a single, specially expressed instance of a relationship is all that is required for this method; and, 3) it can be applied to a wide range of texts. Additionally, the LSP method has the advantage of learning both concepts and relationships at the same time. However, a notable drawback of this method, identified by Hearst, is the low recall.

Many researchers have followed in Hearst’s footsteps, further refining the LSP method for concept and relationship discovery. Mukherjea and Sahay [41] explored how to combine a World Wide Web (WWW) search engine and Hearst patterns for biomedical relationship discovery. The assumption was that if a biological term belongs to a particular class, there should be a large number of Hearst patterns containing that term and that class on the WWW. For example, malaria is a disease, so the phrase “diseases ***including*** malaria” should occur frequently on web pages. They first queried Google Web search engine with hand-crafted LSPs. Then, they used the BioAnnotator system they developed to identify biomedical terms. If the total number of patterns containing the term exceeded a predefined threshold, the term was defined as a member of the class. For evaluation, they randomly selected 100 UMLS terms belonging to 10 classes of UMLS. They achieved 87.5% precision, 70.2% recall and 77.9% F-score when using 25 as the threshold.

Berland [42], Sundblad [43], and Girju [13] extended the LSP method for part-whole relationship discovery. Berland combined the LSP method with statistical methods and applied the hybrid method to a very large corpus. The output of the method was an ordered list of possible parts for a list of six seed whole objects. Berland achieved 55% accuracy.

Fiszman et al. [44] have shown that non-lexically cued appositive pattern can be used to improve SemRep’s overall accuracy by providing more specific semantic predictions. For example, given a sentence “market authorization has been granted in France for *pilocrapine*, an old *parasympathomimetric agent*, in the treatment of *xerostomia*”, the appositive pattern captured the hypernymic position “*Pilocrapine*-ISA-*Parasymphomimetic Agents*”. From this, a more accurate semantic association “*Pilocrapine*-TREATS-*Xerostomia*” over “*Parasympathomimetric Agents*-TREATS-*Xerostomia*” could be inferred. Using LSP increased SemRep’s recall by 7% (39% to 46%) and precision by 1% (77% to 78%).

In this study, we sought to determine the utility of the LSP matching approach for extracting concepts from free-text clinical documents, a rich electronic resource. We first determined the frequency of known LSPs in two large clinical corpora, and then studied the yield of new concepts. We further examined the utility of this method for relationship extraction by

characterizing the types of relationships expressed in each pattern, along with their prevalence. During the course of this research, we also refined a methodology and set of metrics that can be used to estimate the value of various approaches to ontology learning from text.

II. RESEARCH QUESTIONS

1. What is the prevalence and distribution of known LSPs in clinical corpora?
2. What is the value of the LSP matching method for biomedical ontology enrichment using clinical documents?

III. METHODS

An overview of the methods used in this study is provided in flowchart form in Figure 1.

A. Lexico-syntactic patterns (LSPs)

We first identified a set of LSPs for use in this study. The set of LSPs included those identified by Hearst [31] and Berland [42], supplemented by some from our own manual inspection of clinical documents. Table I lists LSPs used in this study and provides example sentences that contain patterns observed in the corpora.

B. Clinical corpora

We used two clinical document types as ontology learning resources - surgical pathology reports and radiology reports. 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 of the University Pittsburgh Medical Center (UPMC), which includes a total of 18 hospitals. Both corpora were de-indentified to meet the requirements of HIPAA “safe harbor” [45]. Use of the clinical corpora was approved by the University of Pittsburgh Institutional Review Board (IRB# PRO07070252).

C. Targeted biomedical knowledge resources

We selected two biomedical knowledge resources in active development that had the potential to benefit from ontology enrichment using clinical text. The National Cancer Institute Thesaurus (NCIT) [46] 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, Biological Process, etc. RadLex [47] is a lexicon for the uniform indexing and retrieval of radiology information resources, sponsored by the Radiology Society of North American (RSNA). It includes over 11,000 concepts in 12 categories, including Imaging Observation, Procedure, Characteristic, Treatment, etc. RadLex has previously been used to derive an application ontology for radiologic reporting, and seems likely to evolve into a formal ontology.

D. Extraction of sentences containing LSPs

Free-text pathology and radiology reports were processed in two steps (Figure 1). First, we tagged Parts-Of-Speech (POS) using a maximum entropy POS tagger that we had previously retrained with pathology reports [48]. Second, we used regular expressions over POS tags to extract all of the LSPs shown in Table I. For example, in the sentence “Compatible with benign eccrine neoplasia, ***such as*** nodular hidroadenoma”, “benign eccrine neoplasia” and “nodular hidroadenoma are Noun Phrases (NPs) and will match the LSP: NP_0 ***such as*** NP_1.

This sentence will be extracted for presentation to the domain experts. The output was a list of all sentences containing LSPs for each corpus. Processing was performed using the GATE platform [49].

E. LSP frequencies and distributions

For each LSP, we calculated the number of documents and sentences that contained the LSP. Because many sentences contained the same terms and LSPs, we also calculated the number of sentences containing unique LSPs. Frequency data enabled us to compute the potential yield of concepts and relationships within a corpus if the rate at which LSPs provide useful information for ontology or lexicon curators is known. Additionally, we used frequency data to determine the sample number for each LSP that was provided to human judges. For LSPs with more than 50 unique instances, we sampled 50 instances. For LSPs with 25 to 50 unique instances, we included all instances. We excluded LSPs with fewer than 25 unique instances.

F. Evaluation of ontology suggestions

We developed a two-step process to determine the value of suggestions generated with the LSP approach. The evaluation approach relies on manual annotations, assuming that automated methods using POS and noun-phrase identification can later be used to approximate the results of the human annotation.

In Step 1, domain experts examined each sentence containing an LSP and identified the Medically Meaningful Terms (MMTs) before and after the LSPs. From the manual annotations, we can evaluate the maximum yield we could expect from applying LSPs to each corpus when we can assume that all the MMTs are correctly extracted. Manually-identified MMTs from Step 1 were used as input to Step 2, in which we evaluated the value of the MMTs for ontology enrichment. The use of human annotations of MMTs, for the evaluation of Step 2, permitted us to more accurately determine the true value of ontology enrichment without confounding the evaluation with possibly incorrect MMTs.

In Step 2, NCIT and RadLex curators determined if the MMT was already present in the knowledge resource and, if not, whether it should be added. Next, they judged whether there was a relationship between the paired MMTs. If there was a relationship, they annotated the type of relationship and indicated if this relationship already existed in the ontology. If it did not exist, they determined if it should be added. Finally, if it should not be added, they provided a reason why it should not be added. We restricted the relationship types to synonym, hypernym, meronym and other (if the relationship does not fall into any of the three pre-determined relationships). These judgments required not only domain knowledge but also in-depth understanding about a knowledge resource's structure and content.

Step 1: Identify the medically meaningful terms from extracted sentences

Domain experts included two resident pathologists (second and third year) and two resident radiologists (second and fourth year). Each group was presented with a sample of LSP-containing sentences from the pathology or radiology corpus, respectively. Domain experts were asked to annotate the MMTs, before and after the LSP, that could stand alone. For example, in the following text "Abnormal slightly high T2 signal seen in the porta hepatis which may be secondary to an underlying malignancy ***such as*** Klatskin tumor or gall bladder carcinoma", the bold and italic term "***such as***" is the LSP. Domain experts would annotate "malignancy" as the MMT before the LSP and "Klatskin tumor" and "gall bladder carcinoma" as the MMTs after the LSP. The final product of the annotation was a table of paired MMTs from each sentence. When multiple terms were annotated before or after the LSP, we created a separate term-pair for each combination. All annotation was performed

using Microsoft Excel. Domain experts were given a spreadsheet containing the sentences extracted with bolded LSPs. They annotated the MMTs before and after the LSP by copying and pasting them into a second and third column.

Domain experts were trained to perform the annotation using a modification of an existing annotation guideline for manual annotation of clinical conditions from emergency department reports developed by Chapman et al [50]. On a development set, we used a Delphi method with repeated training until the F measure exceeded the threshold of 0.9, as depicted in Figure 2. Subsequently, expert annotators were given the final sample to annotate, which consisted of 50 unique sentences for each LSP.

Step 2: Determine the value of concepts and conceptual relationships obtained from MMTs

Domain expert annotations resulted in a list of paired MMTs for pathology and a similar list for radiology. We then invited two experienced curators to judge the MMTs produced by the domain experts in Step 1. One ontology curator, a pathologist who is currently curating the National Cancer Institute Thesaurus, evaluated the term list obtained from the surgical pathology corpus. The other curator, a radiologist who is currently curating RadLex, evaluated the term list obtained from the radiology corpus.

For each term in a term-pair, curators judged:

1) Is the term already represented in the resource (possibly as a synonym)?

2) If not, should a new concept based on this term be added to the resource?

3) If not, what is the reason for which it should not be added?

For each pair of terms, ontology curators also judged:

4) If there is a relationship between the two terms, what is the relationship?

We restricted the choices 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 which it should not be added?

The classical 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 defined more specific evaluation metrics to quantify efficacy for the two discrete steps.

Concept Suggestion Rate (CSR)—

$$\text{CSR}=\frac{\#\text{ of MMTs that were not in the ontology}}{\text{Total \# of MMTs extracted by the enrichment method}} \qquad \text{Equation 1}$$

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

Concept Acceptance Rate (CAR)—

$$CAR=\frac{\#\text{ of MMTs that should be included as new concept, instance, or synonym in the ontology}}{\text{Total \# of MMTs extracted by the enrichment method}} \quad \text{Equation 2}$$

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

Relationship Suggestion Rate (RSR)—

$$RSR=\frac{\#\text{ of relationships that were not in the ontology}}{\text{Total \# of relationships extracted by the enrichment method}} \quad \text{Equation 3}$$

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

Relationship Acceptance rate (RAR)—

$$RAR=\frac{\#\text{ of relationships that should be included in the ontology}}{\text{Total \# of relationships extracted by the enrichment method}} \quad \text{Equation 4}$$

This metric indicates the percentage of concept relationships extracted using the enrichment method that would be added to the relevant ontology.

Additionally, we defined two measures that combine this information to provide an estimate of the total number of concepts or relationships extracted from a given corpus using the LSP matching method.

Estimated Concept Yield (ECY) $_{LSP}$—

$$ECY_{LSP}=N^{*}R^{*}CAR \quad \text{Equation 5}$$

N: Total number of unique LSP in the corpus

R: Average number of MMTs that can be extracted per LSP which is equal to total number of MMTs divided by total number of LSPs

CAR: Concept Acceptance Rate

Estimated Relationship Yield (ERY) $_{LSP}$—

$$ERY_{LSP}=N^{*}P^{*}RAR \quad \text{Equation 6}$$

N: Total number of unique LSP in the corpus

P: Prevalence of a single relationship which is equal to the percentage of a single type of relationship among all of the relationships being extracted.

RAR: Relationship Acceptance Rate

IV. RESULTS

Table II shows the frequency of seven LSPs across the radiology and pathology corpora. Sentences that contained any LSP were extracted. The data are shown as LSPs per sentence and per unique sentence. The overall frequency of patterns appearing in the corpora was low. Although it is not possible to determine how accurate the LSPs are in extracting all relevant instances, the method is expected to perform well in this regard because it is based on string matching. We have not observed false negatives during manual inspection of sample documents from the corpus. Nevertheless, there are factors that could affect the accuracy of results: 1) the POS tagging error; and, 2) it is possible that some instances of the LSP are missed due to misspellings and other typographical errors. The POS tagger was trained with pathology corpus and achieved 93% POS tagging accuracy and the tagging accuracy was 91% when used to tag the radiology reports.

Table III shows the number of medically meaningful terms (MMTs) that could be identified by domain experts in a sample of sentences obtained from each corpus. The total number of sentences used for each LSP is shown in Table II. For each LSP, there was at least one MMT preceding the LSP and more than one MMT following the LSP. Thus, multiple MMTs can be extracted from a sentence that contained a single LSP.

Table IV shows the new concept suggestion rate and the new concept acceptance rate as determined by the curators. For NCIT, the concept suggestion rates ranged from 37% for the pattern "NP such as NP_1, NP_2" to 11% for the pattern "NP of NP_1" with an average of 24% over seven patterns. For RadLex, the suggestion rates were higher, ranging from 52% for the pattern "NP such as NP_1, NP_2" to 18% for the pattern "NP in NP", with an average of 37% over five patterns. However, nearly all the terms suggested would be accepted into the NCIT. The concept suggestion rate and concept acceptance rate were nearly equal. In contrast, for RadLex, the majority of terms suggested would not be accepted into the terminology as judged by the curator.

One of the advantages of the LSP matching method is that the extracted terms preceding and following the LSP are expected to be semantically related. In our study, curators evaluated the semantic relationships between the pairs of MMTs and we calculated the distribution of each type of relationship based on the curator annotation (Table V).

Table VI shows the new relationship suggestion rate and the new relationship acceptance rate as determined by the curators. For NCIT, on average, the relationship suggestion rate was 64%, and the relationship acceptance rate was 14%. For RadLex, on average, the relationship suggestion rate was 55%, and the relationship acceptance rate was 44%.

Using the metrics Estimated Concept Yield (ECY) and Estimated Relationship Yield (ERY) for both pathology corpus and radiology corpus, we estimated that as many as 15,000 (for radiology corpus) to 16,000 (for pathology corpus) new concepts, instances or synonyms could be added, and perhaps as many as 2,000 (for pathology corpus) to 5,000 (for radiology corpus) new relationships could be added.

We also explored reasons why some of the suggested relationships would not be added into the corresponding resource. The top three reasons were: 1) the relationships between classes of concepts are not modeled in the ontology (60%) (e.g. NCIT does not support relationships between anatomic concepts, procedure concepts, and findings); 2) the relationship between two concepts is too general or vague to be included (20%) (e.g. the relationship between "complication" and "Primary biliary cirrhosis" was considered to be too general); and, 3) there is no relationship between the two extracted concepts (10%).

V. DISCUSSION

Our results indicate the Lexico-Syntactic Pattern (LSP) matching method is an effective method for semantic information extraction from clinical documents but also point to some limitations of this approach.

On the positive side, the method can be expected to produce many suggestions for new concepts, instances, synonyms and relationships. Several factors contribute to this finding. First, each instance of a pattern that appeared in the text resulted in extraction of more than two MMTs per sentence. Second, for both corpora tested, at least one quarter of the terms that could be extracted were not associated with corresponding concepts in the existing knowledge resource. With regard to acceptance, the results were mixed. For the pathology corpus, nearly all of these terms were accepted by the curator as useful concepts for the ontology. For the radiology corpus, less than one third of the suggested concepts were accepted by the curator as useful.

In many cases, the scope and structure of the knowledge resource was the limiting factor in concept acceptance. Using this approach, more than half of the relationships identified in the text corpora were not present in either resource. However, curators rated these relationships for acceptance quite differently between NCIT and RadLex, with a much lower overall acceptance rate for NCIT when compared with RadLex. The low relationship acceptance rate for NCIT was mainly due to the fact that many relationships in the text were not within the scope of the ontology. For example, relationships between findings and disease are not defined in the NCIT but these relationships are plentiful in the corpus. In future work, syntactic information derived from concepts and conceptual relationships in the ontology could be used to further constrain suggestions. Selecting candidate concepts based on the type of relationships modeled in the ontology might increase the acceptance rate by limiting suggestions that are clearly not modeled in the ontology.

The value of the LSP matching method also depends on how frequently these patterns exist in a domain corpus given that these patterns are likely to extract more meaningful medical terms. However, quantity is not the only measure. Our study showed that distributions of LSPs are heterogeneous. Some LSPs have higher frequencies than others and these proportions differed across the two corpora we studied. Importantly, some of the patterns can be highly effective because a single specifically expressed instance of a relationship was all that was required for new semantic knowledge extraction. For example, even though “NP_aka_NP” was a low frequency pattern, from a single instance of the “NP_aka_NP”, “Schwannoma (aka neurilemoma)” we can extract a correct synonym relationship between schwannoma and neurilemoma. The frequencies of patterns in two different types of clinical documents were different and some of the patterns either did not exist or had very low frequency (e.g. “NP_aka_NP” and “NP_so called_NP” in radiology reports). Furthermore, the frequency of the patterns in the corpus does not guarantee the high suggestion rate since the top three patterns based on suggestion rates are “NP such as NP_1, NP_2”, “NP including NP_1, NP_2”, and “NP other NP_1, NP_2” for both corpora. Yet these three patterns have relatively low frequencies in both corpora.

To determine the overall value of the LSP pattern set as a method of semantic extraction, we computed an estimate of the yield of concepts and relationships for each corpus. For a large corpus, the yield of concepts and relationships could be quite substantial.

On the negative side, in contrast to Hearst’s paper, our results suggest that there is little information in the LSP that can accurately predict the semantic relationship between the concepts in the LSP, at least in the two domains studied. The distribution of relationships extracted with the LSP method was heterogeneous. Of the three named relationship types

that curators evaluated (hyponym, meronym, and synonym), the most frequent relationship extracted with “NP_such as_NP”, “NP_including_NP” or “NP_other_NP” patterns was hypernym/hyponym, and the most frequent relationship extracted using “NP_aka_NP’ was synonym. In many cases, the relationship was determined to be of some other type. In some cases, there was no identifiable relationship between the Meaningful Medical Terms extracted. Thus, we will not be able to use the LSP as an indicator of the type of relationship expressed between the entities. However, because LSP extracts terms in pairs, if one of the terms extracted by the LSP method is already in the ontology, it could be informative in determining a general position in the hierarchy for a concept based on the complementary term. This could be a very useful feature for a semi-automated ontology learning platform where a human curator is required to determine the type of relationship between two terms.

Another limitation of this method, pointed out by previous research, is low recall. It is likely that many candidate concepts in the corpus never appear in any pattern. Attempts to improve the recall of the LSP method have focused on three major approaches. The first approach is to use additional syntactic features such as noun coordination information in combination with LSPs. For example, consider the following sentence containing a coordination structure: “In the ovine brain, GnRH neurons do not contain type II glucocorticoid (GR), progesterone (PR), or α estrogen (ERα) receptors”. If “ERα” is a steroid receptor in the ontology, the assumption that coordinated concepts are related, permits defining “GR” and “PR” also as steroid receptors. Caraballo [28] and Cederberg [29] used this approach to obtain additional related pairs of terms. The second approach is to use machine learning method by learning new patterns with either seed terms Riloff [51], Downey [30] or seed patterns Xu [52] in an iterative bootstrapping process. Finally, a third approach combines pattern and co-occurrence information to learn new patterns. As an example of this approach, Pantel et al. [53] used the minimal edit distance algorithm for pattern learning.

A final limitation of the LSP method is that it focuses on the use of simple English patterns rather than domain specific patterns. The pattern learning approaches discussed above have been applied to specific corpora to learn domain specific extraction patterns [51,54]. For example, Embarek and Ferret [55] discovered many medically related patterns using Pantel’s algorithm and used these patterns to discover semantic relationships in the medical domain. Patterns discovered this way showed good results when they were evaluated using a medical corpus of the EQueR evaluation campaign for question-answering systems in French. We believe that future enhancements that build on the work of Riloff [51], Snow [54], and other investigators [28,29] could reduce the limitations of the LSP method for ontology enrichment in biomedicine.

In summary, we conclude that Lexico-Syntactic Pattern matching is an effective, but limited, method for concept and relationship discovery in clinical corpora. We view this method as complementary to other enrichment methods. It seems likely that approaches combining multiple methods of concept discovery (including LSP) will prove most useful for semi-automated enrichment of biomedical ontologies.

VI. LIMITATIONS

One limitation of our study was that we did not specifically determine the percentage of MMTs that represented synonyms of existing ontology concepts as opposed to unrecognized new concepts. This would be an important extension of our current findings. It would help us to separate out the proportion of extracted elements that provide information for the various ontology learning tasks (e.g. synonym, concept and relationship extraction). In designing the present study, we elected to exclude this determination. We reasoned that the human annotation of MMTs essentially eliminated uninformative extraction. Future

evaluations of automated LSP extraction will provide a more appropriate setting for assessing the relative value of LSP matching across the ontology learning tasks.

Acknowledgments

The authors wish to thank Karma Lisa Edwards and Lucy Cafeo of the University of Pittsburgh for editorial assistance, and Kevin Mitchell for expert technical help. We thank Dr. David Chanin for his help in measuring the LSP matching method for enrichment of RadLex. This work was supported by NIH grant RO1 CA 127979.

REFERENCES

1. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucl Acids Res. 2004; 32 suppl_1:267–270.
2. Cowell, L.; Smith, B. Infectious disease ontology. In: Sintchenko, V., editor. Infectious Disease Informatics. Springer: New York City; 2010. p. 373-395.
3. The Gene Ontology Consortium. The Gene Ontology project. Nucl Acids Res. 2008; 36 suppl_1:440–444.
4. HL7: HL7 Reference Information Model. Available from: http://www.hl7.org/implement/standards/rim.cfm.
5. Achour S, Dojat M, Rieux C, Bierling P, Lepage E. A UMLS-based Knowledge Acquisition Tool for Rule-based Clinical Decision Support System Development. J Am Med Inform Assoc. 2001; 8(4):351–360. [PubMed: 11418542]
6. Collier N, Kawazoe A, Jin L, Shigematsu M, Dien D, Barrero RA, et al. A multilingual ontology for infectious disease surveillance: rationale, design and challenges. Language Resources and Evaluation. 2006; (40):405–413.
7. Kashyap, V.; Morales, A.; Hongsermeier, T. On implementing clinical decision support: achieving scalability and maintainability by combining business rules and ontologies. Proceedings of the Annual Symposium of American Medical Informatics Association; Washington, DC. 2006. p. 414-418.
8. Haynes B, McKibbon A, Wilczynski N, Walter S, Werre S. for the Hedges T. Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. BRIT MED J. 2005; 330(7501):1179. [PubMed: 15894554]
9. Sneiderman CA, Demner-Fushman D, Marcelo Fiszman M PhD, Ide NC, Rindflesch TC. Knowledge-based methods to help clinicians find answers in MEDLINE. J Am Med Inform Assoc. 2007; 14(6):772–780. [PubMed: 17712086]
10. Meystre S, Haug PJ. Natural language processing to extract medical problems from electronic clinical documents: Performance evaluation. J Biomed Inform. 2006; 39(6):589–599. [PubMed: 16359928]
11. Liang, T.; Lin, Y-H. Anaphora Resolution for Biomedical Literature by Exploiting Multiple Resources. In: Dale, R.; Wong, K-F.; Su, J.; Kwong, OY., editors. Natural Language Processing – IJCNLP. Berlin / Heidelberg: Springer; 2005. p. 742-753.
12. Pustejovsky, J.; Rumshisky, A.; Castano, J. Language Resources and Evaluation Workshop on Ontologies and Lexical Knowledge Bases. Spain: Las Palmas, Canary Islands; 2002. Rerendering semantic ontologies: Automatic extensions to UMLS through corpus analytics; p. 60-67.
13. Girju, R.; Badulescu, A.; Moldovan, D. Learning semantic constraints for the automatic discovery of part-whole relations. Proceedings of the Human Language Technology Conference; Canada: Edmonton; 2003. p. 80-87.
14. Wagner C. End-users as expert system developers. Journal of End User Computing. 2000; 12(3):3–13.
15. Wagner C. Breaking the knowledge acquisition bottleneck through conversational knowledge management. Information Resources Management. 2006; 19(1):70–83.
16. Waterman, DA. A guide to expert systems. Addison-Wesley Longman Publishing Co., Inc.; 1985.
17. Druss BG, Marcus SC. Growth and decentralization of the medical literature: implications for evidence-based medicine. J Med Libr Assoc. 2005; 93(4):499–501. [PubMed: 16239948]

18. Chun, H-W.; Tsuruoka, Y.; Kim, J-D.; Shiba, R.; Nagata, N.; Hishiki, T. Extraction of gene-disease relations from Medline using domain dictionaries and machine learning. Proceedings of Pacific Symposium on Biocomputing; Maui, HI: 2006. p. 4-15.

19. Collier, N.; Park, H.; Ogata, N.; Tateishi, Y.; Nobata, C.; Ohta, T., et al. The GENIA project: corpus-based knowledge acquisition and information extraction from genome research papers. 9th Conference of the European Chapter of the Association for Computational Linguistics; Norway: Bergen; 1999. p. 271-272.

20. Daraselia N, Yuryev A, Egorov S, Novichkova S, Nikitin A, Mazo I. Extracting human protein interactions from MEDLINE using a full-sentence parser. Bioinformatics. 2004; 20(5):604–611. [PubMed: 15033866]

21. Chapman WW, Dowling JN, Wagner MM. Fever detection from free-text clinical records for biosurveillance. J Biomed Inform. 2004; 37(2):120–127. [PubMed: 15120658]

22. South, BR.; Chapman, WW.; Delisle, S.; Shen, S.; Kalp, E.; Perl, T., et al. Optimizing A Syndromic Surveillance Text Classifier for Influenza-like Illness: Does Document Source Matter?. Proceedings of the Annual Symposium of American Medical Informatics Association; Washington, DC. 2008. p. 692-696.

23. Cornet R, De Keizer NF, Abu-Hanna A. A framework for characterizing terminological systems. Methods Inf Med. 2006; 45:253–266. [PubMed: 16685333]

24. de Keizer NF, Abu-Hanna A. Understanding terminological systems II: terminology and typology. Methods Inf Med. 2000; 39:22–29. [PubMed: 10786066]

25. de Keizer NF, Abu-Hanna A, Zwetsloot-Schonl JHM. Understanding terminological systems I: terminology and typology. Methods Inf Med. 2000; 39:16–21. [PubMed: 10786065]

26. Buitelaar, P.; Cimiano, P.; Magnini, B. Ontology learning from text: method, evaluation and applications. Breuker, J.; Dieng, R.; Guarino, N.; Mantaras, RLd; Mizoguchi, R.; Musen, M., editors. Amsterdam, Berlin, Oxford, Tokyo, Washington DC: IOS Press; 2005.

27. Gomez-Perez A, Manzano-Macho D. An overview of method and tools for ontology learning from texts. The Knowledge Engineering Review. 2005; 19(3):187–212.

28. Caraballo, S. Automatic construction of a hypernym-labeled noun hierarchy from text. Proceedings of the 37th Conference on Computational Linguistics; College Park, MD. 1999. p. 120-126.

29. Cederberg, S.; Widdows, D. Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. Proceedings of the 7th Conference on Natural Language Learning; Canada: Edmonton; 2003. p. 111-118.

30. Downey, D.; Etzioni, O.; Soderland, S.; Weld, DS. Proceedings of the American Association for Artificial Intelligence Workshop on Adaptive Text Extraction and Mining. San Jose, CA: 2004. Learning text patterns for Web information extraction and assessment; p. 50-55.

31. Hearst, MA. Automatic acquisition of hyponyms from large text corpora. Proceedings of the 12th Conference on Computational Linguistics; France: Nantes; 1992. p. 539-545.

32. Church, KW.; Hanks, P. Word association norms, mutual information, and lexicography. Proceedings of 27th Annual Meeting of the Association for Computational Linguistics; Vancouver, BC, Canada. 1989. p. 76-83.

33. Grefenstette, G. Sextant: exploring unexplored contexts for semantic extraction from syntactic analysis. Proceedings of the 30th annual meeting of the Association for Computational Linguistics; Newark, DE. 1992. p. 324-326.

34. Grefenstette, G. Explorations in automatic thesaurus discovery. Boston, MA: Kluwer Academic Publisher; 1994.

35. 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.

36. Nenadć, G.; Spasić, I.; Ananiadou, S. Proceedings of the 2nd International Workshop on Computational Terminology. Taipei, Taiwan: Association for Computational Linguistics; 2002. Automatic discovery of term similarities using pattern mining; p. 1-7.

37. Ryu, P-M.; Choi, K-S. Measuring the specificity of terms for automatic hierarchy construction. Proceedings of European Conference on Artificial Intelligence Workshop on Ontology Learning and Population; Valencia, Spain. 2004.

38. Liu K, Hogan WR, Crowley RS. Natural language processing methods and systems for biomedical ontology learning. J Biomed Inform. 2010 In press.
39. ODIE toolkit. 2010 Available from: http://bioontology.org/tools/ODIE.html.
40. Crowley, RS.; Chavan, G.; Mitchell, K.; Liu, K.; Savova, G.; Chapman, W., et al. Proceedings of Annual Symp of American Medical Informatics Association. Washington, DC: 2010. ODIE – A workbench for cyclic entity recognition and ontology enrichment. Submitted
41. Mukherjea, S.; Sahay, S. Discovering biomedical relations utilizing the World-Wide Web. Proceedings of Pacific Symposium on Biocomputing; Maui, HI. 2006. p. 164-175.
42. Berland, M.; Charniak, E. Finding parts in very large corpora. Proceedings of the 37th Conference on Computational Linguistics; College Park, MD. 1999. p. 57-64.
43. Sundblad, H. Automatic acquisition of hyponyms and meronyms from question corpora. Proceedings of the 15th European Conference on Artificial Intelligence; Lyon, France. 2002.
44. Fiszman, M.; Rindflesch, TC.; Kilicoglu, H. Integrating a hypernymic proposition interpreter into a semantic processor for biomedical texts. Proceedings of the Annual Symposium of American Medical Informatics Association; Washington, DC. 2003. p. 239-243.
45. Health Insurance Portability and Accountability Act of 1996. Available from: http://aspe.hhs.gov/admnsimp/pl104191.htm
46. National Cancer Institute Thesaurus (NCIT). 2010. Available from: http://ncit.nci.nih.gov/
47. 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
48. Liu K, Chapman W, Hwa R, Crowley RS. Heuristic Sample Selection to Minimize Reference Standard Training Set for a Part-Of-Speech Tagger. J Am Med Inform Assoc. 2007 September 1; 14(5):641–650. 2007. [PubMed: 17600099]
49. GATE. 2010 June. Available from: http://gate.ac.uk/.
50. Chapman WW, Dowling JN, Hripcsak G. Evaluation of training with an annotation schema for manual annotation of clinical conditions from emergency department reports. Int J Med Inform. 2008; 77(2):107–113. [PubMed: 17317291]
51. Riloff, E. Automatically generating extraction patterns from untagged text. Proceedings of the 13th National Conference on Artificial Intelligence; Portland, OR. 1996. p. 1044-1049.
52. Xu, R.; Morgan, A.; Das, AK.; Garber, A. Proceedings of the Workshop on BioNLP. Boulder, Colorado: 2009. Investigation of unsupervised pattern learning techniques for bootstrap construction of a medical treatment lexicon; p. 63-70.
53. Pantel, P.; Ravich, D.; Hovy, E. Towards terascale knowledge acquisition. Proceedings of Conference on Computational Linguistics; Barcelona, Spain. 2004. p. 771-777.
54. Snow, R.; Jurafsky, D.; Ng, AY., editors. Learning syntactic patterns for automatic hypernym discovery. Cambridge, MA: MIT Press; 2005.
55. Embarek, M.; Ferret, O. Learning patterns for building resources about semantic relations in the medical domain. Proceedings of the 6th International Language Resources and Evaluation; Marrakech, Morocco. 2008. p. 2006-2012.

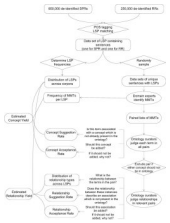

Fig. 1.

Overall study methodology

The sizes of corpora in this figure have been rounded.

SPRs: Surgical Pathology Reports, RRs: Radiology Reports, POS: Part-of-Speech, LSP: Lexico-Syntactic-Pattern, MMTs: Medically Meaningful Terms

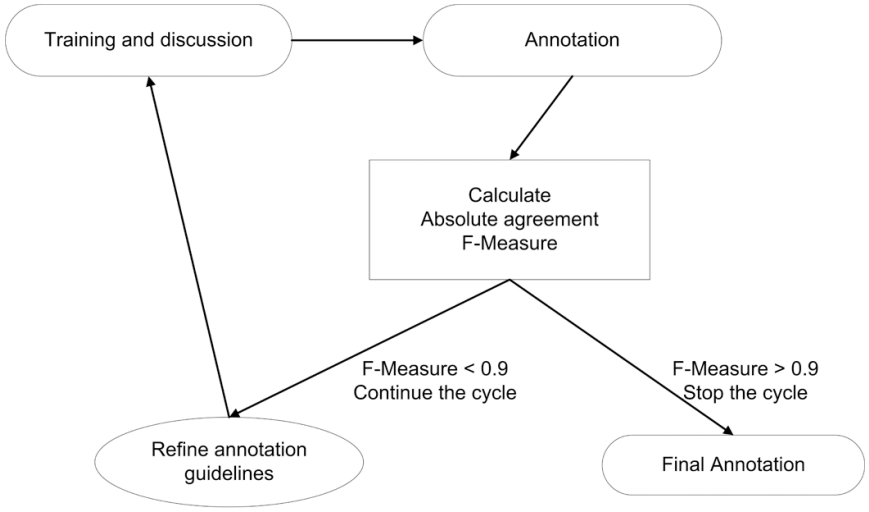

Fig. 2.
Domain expert training process

Table I

Lexico-Syntactic Patterns with examples from corpora

LSP Category	LSP	Examples
Hearst	NP_0 such as {NP_1, NP_2 ..., and|or} NP_n Such NP_0 as {NP_1,}* {or | and} NP_n	Compatible with benign eccrine neoplasia, **such as** nodular hidroadenoma
		Such atypical pneumonia **as** mycoplasma or viral pneumonitis
	NP_1 {, NP_2} * {,} or other NP_0 NP_0{, NP_1}*{,} and other NP_2	Residual basal cell carcinoma **or other** malignancy
		Pneumoconiosis **and other** chronic process
	NP_0{,} including {NP_1,}*{or | and} NP_2	Peripheral blood pancytopenia **including** macrocytic anemia and rare nucleated red blood cells
		Chronic obstructive pulmonary disease **including** bronchial wall thickening
Other	NP_0 [a.k.a.|aka | also known as] NP_1 *{or | and} NP_n	Sebaceoma (**aka** sebaceous epithelioma)
	NP_0 so called {NP_1,}*{or | and} NP_n	Pleomorphic adenoma (**so called** hybrid adenoma)
Berland	Part NN in PREP {the | a} DET mods [JJ|NN]* whole NN Parts NN-PL in PREP wholes NN-PL	Phospholipids **in** the cell membrane…
	Part NN-PL of PREP {the | a} DET mods [JJ|NN]* whole NN Parts NN-PL of PREP wholes NN-PL	Membrane **of** a cell

NP: Noun Phrase, NN: Noun, PREP: preposition, DET: determiner, JJ: adjective, NN-PL: Noun plural form, mods: modifiers. Regular expression notation: {x}: x is optional, x|y: either x or y, x*: zero or more of repetition of x

Table II

Frequencies of various LSPs in the pathology and radiology corpora

LSP	Surgical pathology reports			Radiology reports		
	852,764 reports, 16,157,608 sentences			209,997reports, 4,057,228 sentences		
	# sentences	# sentences containing unique LSP	# randomly selected sentences	# sentences	# sentences containing unique LSP	# randomly selected sentences
NP such as NP	98	95	50	906	251	50
NP including NP	6291	4952	50	1403	747	50
NP other NP	6940	2251	50	10622	1407	50
NP also called NP	48	37	37	29	22	0
NP aka NP	5396	460	50	2	2	0
NP in the NP	47124	23178	50	64044	29285	50
NP of the NP	246798	70735	50	173016	54895	50
total	312695	101708	337	250022	86609	250

Table III

Number of medically meaningful terms (MMTs) extracted by the LSP method

LSP	Surgical pathology reports		Radiology reports	
	Preceding the LSP	Following the LSP	Preceding the LSP	Following the LSP
	Ratio (# of MMTs/ # of instances of LSP)	Ratio (# of MMTs/ # of instances of LSP)	Ratio (# of MMTs/ # of instances of LSP)	Ratio (# of MMTs/ # of instances of LSP)
NP such as NP_1, NP_2	1.04 (52/50)	1.88 (94/50)	1.0 (50/50)	1.9 (95/50)
NP including NP_1, NP_2	0.98 (49/50)	1.62 (81/50)	1.0 (50/50)	1.72 (86/50)
NP other NP_1, NP_2	1.0 (50/50)	1.06 (53/50)	1.0 (43/43)	1.0 (43/43)
NP also called NP_1, NP_2	0.95 (35/37)	0.97 (36/37)	NA	NA
NP aka NP_1, NP_2	0.96 (47/50)	1.18 (59/50)	NA	NA
NP in NP_1	1.0 (50/50)	1.0 (50/50)	0.94 (47/50)	0.76 (39/50)
NP of NP_1	1.0 (50/50)	1.0 (50/50)	0.8 (40/50)	0.68 (34/50)
Average	0.99 (333/337)	1.26 (423/337)	0.95 (230/243)	1.22 (296/243)
Average # MMT per LSP	2.25		2.21	

Table IV

Comparison of new concept suggestion rate and acceptance rate

LSP	Surgical pathology reports		Radiology reports	
	CSR	CAR	CSR	CAR
NP such as NP_1, NP_2	37% (52/140)	31% (43/140)	52% (75/145)	10% (14/145)
NP including NP_1, NP_2	32% (61/189)	32% (60/189)	39% (54/138)	14% (19/138)
NP other NP_1, NP_2	16% (18/113)	16% (18/113)	33% (28/86)	8% (7/86)
NP also called NP_1, NP_2	14% (10/74)	10% (7/74)	NA	NA
NP aka NP_1, NP_2	31% (37/119)	31% (37/119)	NA	NA
NP in NP_1	12% (12/100)	6% (6/100)	18% (13/74)	8% (6/74)
NP of NP_1	11% (11/98)	6% (6/98)	26% (21/80)	14% (11/80)
Average	24% (201/833)	21%(177/833)	37% (191/523)	11% (57/523)

Table V

Distribution of semantic relationships extracted using the LSP matching method

Corpus	LSP	Semantic Relationship				
		Hyponym	**Synonym**	**Meronym**	**Other**	**None**
Surgical Pathology Reports	NP such as NP_1, NP_2	37% (24/65)	0%	2% (1/65)	57% (37/65)	5% (3/65)
	NP including NP_1, NP_2	10% (11/114)	1% (1/114)	6% (7/114)	78% (89/114)	5% (6/114)
	NP other NP_1, NP_2	39% (24/61)	0% (0/61)	2% (1/61)	46% (28/61)	8% (5/61)
	NP also called NP_1, NP_2	22% (9/41)	20% (8/41)	0%	37% (15/41)	10% (4/41)
	NP aka NP_1, NP_2	10% (6/59)	44% (26/59)	0%	39% (23/59)	5% (3/59)
	NP in NP_1	0%	0%	0%	100% (45/45)	0%
	NP of NP_1	5% (2/44)	0%	18% (8/44)	61% (27/44)	16% (7/44)
	Average	18% (76/429)	8% (35/429)	4% (17/429)	62% (264/429)	7% (28/429)
Radiology Reports	NP such as NP_1, NP_2	72% (26/36)	0%	0%	28% (10/36)	0%
	NP including NP_1, NP_2	39% (7/18)	0%	11% (2/19)	33% (6/18)	17% (3/18)
	NP other NP_1, NP_2	76% (16/21)	0%	0%	0%	24% (5/21)
	NP in NP_1	4% (1/26)	0%	12% (3/26)	42% (11/26)	42% (11/26)
	NP of NP_1	4% (4/27)	0%	0%	0%	96% (26/27)
	Average	40% (51/128)	0%	4% (5/128)	21% (27/128)	35% (45/128)

Table VI

Comparison of new concept relationship suggestion rate and acceptance rate

	Pathology reports (Enrich NCIT)		Radiology reports (Enrich RADLex)	
LSP	**RSR**	**RAR**	**RSR**	**RAR**
NP such as NP_1, NP_2	55% (36/65)	26% (17/65)	94% (34/36)	94% (34/36)
NP including NP_1, NP_2	78% (89/114)	15% (17/114)	61% (11/18)	39% (7/18)
NP other NP_1, NP_2	51% (31/61)	8% (5/61)	57% (12/21)	57% (12/21)
NP also called NP_1, NP_2	29% (12/41)	10% (4/41)	NA	NA
NP aka NP_1, NP_2	73% (43/59)	24% (14/59)	NA	NA
NP in NP_1	84% (38/45)	0% (0/45)	50% (13/26)	12% (3/26)
NP of NP_1	64% (28/44)	5% (2/44)	0% (0/27)	0% (0/27)
Average	64% (277/429)	14% (59/429)	55% (70/128)	44% (56/128)