# Lexical Syntactic Patterns and Novel Statistical Measures based Bootstrapping Approach for Evolution of Biomedical Ontologies

B. Sathiya
Data Scientist
SPi Technologies Pvt. Ltd.
Chennai, India

T. V. Geetha
Professor, DCSE
Anna University, Chennai, India

## ABSTRACT

Knowledge extraction and information processing from the proliferating biomedical data is a primary challenge to the researchers in this field. This is tackled by a semantic knowledge representation model with controlled vocabulary termed as ontology. However, the exponential growth of biomedical data makes the ontology outdated soon and hence its evolution process becomes an inevitable one. Even though numerous ontology evolution systems attempted to evolve the ontology automatically in numerous ways, identifying concepts of ontology that need to be evolved and discovery of new components of the concepts such as its related new concepts and relations is not handled automatically. Therefore, the aim of this work is to automatically identify the concepts which need to be evolved and discover the new components for those concepts using the web pages and MEDLINE database. Particularly, a new concept selection measure: CE (**C**oncept to be **E**volved) is designed to select the concepts with high possibility to be evolved based on the number of neighbour and depth of it. Next, a lexical syntactic pattern based bootstrapping approach with new statistical scoring measures such as HH-CS (**H**yponym **H**ypernym-**C**oncept **S**coring), DR-CS CS (**D**omain **R**ange-**C**oncept **S**coring) and RS (**R**elation **S**coring) is proposed to discover new candidate components from web pages using the set of patterns and precisely select the correct candidate components from the MEDLINE database using the scoring measures. The experimental results on the biomedical ontologies in terms of precision, recall, F-measure and ontology quality metrics prove the effectiveness of the proposed CE measure and bootstrapping approach with new statistical measures in precisely identifying concepts to be evolved and discovering new components.

## Keywords

Ontology evolution, enrichment, bootstrapping, biomedical ontologies.

## 1. INTRODUCTION

In the booming era of biomedicine with vast assorted information, sharing and reusing valuable information by tackling the heterogeneity is the prime need of the day. To store, share, reuse, maintain and evolve the information, a proper, well structured, rich and semantic representation is provided by the ontologies. The process of construction ontology from the raw texts is called ontology learning from text. This process is under the assumption that, the raw texts consist of all required information to represent the domain completely. However, in the field of biomedicine, this assumption is tough to be satisfied since the raw texts become outdated as the domain of interest evolves continuously and drastically. Consequently, the biomedical ontologies also become outdated. However, in the recent years, numerous advancements made in the biomedical informatics were through the usage of biomedical ontologies [1-3]. Therefore, it is important to maintain the completeness and freshness of the domain represented by the ontology and hence it needs to be evolved. This leads to the process of ontology evolution which deals with the addition, modification and deletion of the components of ontology such as the concepts, properties, relations and axioms.

Even though numerous methods exist in the literature, there are still many open challenges to be handled by the researchers. One of the foremost open challenges [4, 5] is the identification of the new changes required, i.e. discovery of new concepts and relations between the concepts and positioning it in the existing ontology. According to Khattak et al. [4, 5], most of the existing systems [6- 9] have manually identified the required new changes. Further, systems in [10, 11] uses domain experts and system in [12] employs users to validate the update of the new concepts. Meanwhile, few systems [13-15] had identified the new capture changes between the different versions of the ontology by OntoView [16], PromptDiff (Protege plug-in), and H-Match [17] algorithms. Other systems [18, 11, 19] used generic and limited knowledge base such as WordNet [20], and UNL Knowledge Base and ontology [21] to detect new capture changes in the existing ontology.

Henceforth, full automation of new changes in the evolution process is still a challenge to be handled. From the review, to the best of our knowledge, none of the systems have identified the need for evolution at the concept level (i.e. fine grain level). Further the abundant knowledge available in the web and the large biomedical literature (MEDLINE) is not utilized automatically to effectively evolve the concept indentified. However, both these are issues of the day and need to be addressed since, the biomedical field is getting diversified, intricate and continuously growing in a faster manner. Hence evolution of each concept in the biomedical ontologies is necessary and can be made possible by utilizing the proliferating amount of knowledge available from the ever growing literatures. Hence, in this paper a novel concept selection measure: CE and a new semi-supervised machine learning approach (Bootstrapping) is proposed to identify and evolve the existing ontology using the web and MEDLINE.

The main contributions of this work are as follows. First, the set of concepts which have need or scope to be evolved is identified using a new concept selection measure called CE. Second, a new lexical syntactic patterns based bootstrapping approach is proposed for automatic discovery of new components for the selected concepts from the web pages. Third, a set of novel statistical scoring measures such as HH-CS DR-CS and RS have been proposed to validate the discovery of new components based on the biomedical literature in the proposed bootstrapping approach. The latest biomedical literature required for the validation of new components is obtained from the MEDLINE (https://www.nlm.nih.gov/bsd/pmresources.html) repository consisting of 16 million journal articles, and it is added with 2,000 to 4,000 new articles each day [22]. It is the vast and most commonly used repository for biomedical knowledge extraction [23-26]

The rest of the paper is organized as follows. Section 2 briefs the related works. Section 3 describes the proposed novel and automated ontology evolution process consisting of the CE measure and bootstrapped method in detail. Section 4 outlines the experimental setup and illustrates the various experimental results. Section 5 concludes the paper with a note on future work.

# 2. METHODOLOGY

This research proposes a new measure CE to select the concepts that need to be evolved and novel pattern based bootstrapping approach to evolve the chosen concepts in the existing ontology. First, the concepts to be evolved are chosen using the proposed CE measure. Then, this work aims at identifying new components of the chosen existing concepts of the ontology using the bootstrapping approach. A chosen concept c can be evolved by adding one or more of the following components to it: i) One or more new parent concept(s) termed as hypernym(s)) related to c with the is-a (taxonomical) relationship, (ii) One or more new child concept(s) termed as hyponym(s)) related to c using the is-a (taxonomical) relationship. For example: In MeSH (http://www.nlm.nih.gov/mesh), 'Typhoid Fever' and 'Paratyphoid Fever' are the hyponym concept connected with is-a relation to hypernym concept: 'Enterobacteriaceae Infections' , (iii) One or more new domain concept(s) related to c using the non is-a (non-taxonomical) relation R, (iv) One or more new range concept(s) related to c using the non is-a (non-taxonomical) relation R. For example: 'Disease' is the domain concept linked to the range concept 'Medicine' using the relation R: 'treated-by' and (v) One or more new taxonomical or non-taxonomical relation(s) between existing two concepts c and c'.

The overview of the proposed bootstrapping approach is as follows. First, a set of seed patterns based on the chosen concepts are formulated. Then, using the web resources new patterns are formed from the seed patterns and each new pattern is scored using the proposed scoring measures based on MEDLINE. Further, the patterns are sorted based on scoring and patterns with scoring above the threshold β are chosen. The new concepts and relations in the chosen patterns that clear the redundancy and inconsistency check are used to form seed patterns for the next iteration. This process repeats until no new concepts or relations are discovered in the iteration. In the following sections, the proposed concept selection and bootstrapping approach are explained in detail.

## 2.1 Novel concept selection measure: CE

The novel concept selection measure CE of a concept c is based on the number of neighbours (hypernyms, hyponyms, domain and range concepts of c) and depth of the c. The concept with less number of neighbours and depth possess a good possibility to be evolved. Because, as the depth of the concept increases, its specificity too increases and hence finding more neighbours with distinguishing characteristics is less probable. The $\mathrm{CE(c)} \to \Re \in [0,1)$ is formally defined as follows.

$$CE(c)=\frac{1}{2}\left(1-\frac{\log(neighbour(c))}{\log(\max\_\mathrm{neighbour})}\right)+\frac{1}{2}\left(1-\frac{\log(depth(c))}{\log(\max\_depth)}\right) \quad (1)$$

where, neighbour(c) denotes the number of hypernyms, hyponyms, domain and range concepts of c, depth(c) denotes the depth of concept c in the ontology, max_neighbour represents the number of neighbour of the root concept of the ontology and max_depth corresponds to the maximum depth of the ontology. A larger value of CE(c) indicates that the need or the possibility for c to be evolved is more and vice versa. Based on the score, the concepts are sorted in descending order and top K concepts are chosen as candidate concepts to be evolved using the proposed bootstrapping approach.

## 2.2 Proposed bootstrapping approach

### *2.2.1 Seed pattern formation*

For each candidate concept c, a set of seed pattern is formed to check for the availability of new above mentioned components based on the evidence from the web. The seed pattern is designed to consists of ontology concept to be evolved, key words or non-taxonomical relation label and a wildcard. The structure of the seed pattern differs based on the purpose of it. The structures of the seed patterns to discover new hyponym for a concept c are as follows: (i) < *, Key-Set1, c > and (ii) < c, Key-Set2, * >. Here the sets of keywords: Key-Set1: {and other, or other, is a} and Key-Set2: {such as, including, especially, called, particularly, for example, among which} are obtained from Hearst patterns [27] and snow et al. patterns [28] (Table I). These set of patterns [27, 28] are most effective and predominantly used by numerous systems to identify the hyponyms and hypernyms in the literature. A set of 10 hyponym seed patterns is formed with each of these keywords. For example <*, is a 'Enterobacteriaceae Infections' > is a seed pattern. The wildcard * is a place holder for the new hyponym to be discovered.

Similarly, the structure of the seed pattern to discover new hypernym for a concept c is as follows: (i) < c, Key-Set1, * > and (ii) < *, Key-Set2, c >. The same set of keywords of Hearst pattern [27] and snow et al. pattern [28] is also used here and a set of 10 hypernym seed patterns is formed with each of these keywords. Correspondingly, the structure of the domain and range seed pattern to discover new domain or range concept for a concept c with relation R is as follows: (i) < *, R, c > and (ii) < c, R, * >. Finally, to discover new relations among the existing concepts c and c', the structure of the relation seed pattern is defined as follows: < c, *, c' >. The wildcard in the above relation seed pattern will be filled by keywords (Table I) or a non-taxonomical relation R based on the evidence from the web. In consolidation, for each concept c, 10 hypernym seed patterns, 10 hyponym seed patterns and 1 relation seed pattern is formed. Further, 1 domain seed pattern and 1 range seed pattern is formed provided that there exist a non-taxonomical relation R with c.

**Table 1. The Hearst [27] and Snow et al. [28] patterns**

| Key-Set1 | Key-Set2 |
|---|---|
| HYPONYM, and other HYPERNYM [27] | HYPERNYM, such as HYPONYM [27] |
| HYPONYM, or other | HYPERNYM, including |

| | |
|---|---|
| HYPERNYM [27] | HYPONYM [27] |
| HYPONYM is a HYPERNYM [27] | HYPERNYM, especially HYPONYM [27] |
| | HYPERNYM, called HYPONYM [28] |
| | HYPERNYM, particularly HYPONYM [28] |
| | HYPERNYM, for example HYPONYM [28] |
| | HYPERNYM, among which HYPONYM [28] |

## 2.3 Discovery of new components

To discover the new components, the placeholder: wildcard of seed patterns should be filled by the right candidates based on the evidence from the web pages. To retrieve this information, seed patterns are sent as web queries to the search engine. The web pages are chosen as the knowledge source/evidence since it possesses a huge collection of textual information with the latest trending or emerged concepts. The resultant web pages are searched for the sentences which match the seed patterns. These set of sentences are retrieved, POS tagged and the nouns in the place of wildcard are retrieved as the candidate wildcard fillers. Similar to [29, 30], the multi-term nouns are extracted from the POS-tagged sentences using the following regular expression: $(DT)^{?}(JJ/JJR/JJS)^{*}(NN/NNS/NNP/NNPS)^{+}$. Here DT represents an article, JJ, JJR and JJS denote the adjectives and NN, NNS, NNP and NNPS represent the nouns. This candidate wildcard filler can be a hyponym, hypernym, domain concept, range concept, taxonomical or domain specific relation. Now, each of the seed patterns with each of the newly discovered candidate wildcard fillers forms a new pattern. For example, the hyponym seed patterns <*, 'is a' , 'Enterobacteriaceae Infections' > is given as the web query and the candidate wildcard fillers such as 'Typhoid Fever' and 'Paratyphoid Fever' are retrieved from the web pages.

Now the new hyponym patterns formed are: <'Typhoid Fever', 'is a', 'Enterobacteriaceae Infections' > and <'Paratyphoid Fever', 'is a' , 'Enterobacteriaceae Infections' >. However, these set of newly discovered patterns should be checked for its correctness since the textual information obtained from the web are uncertain and generic. Hence the candidate wildcard fillers in the new patterns are validated using a set of novel scoring measures as follows.

## 2.4 Scoring

Even though numerous statistical measures based on the co-occurrence frequency exist in the literature, they are too generic to detect that there just exist a relation between the two concepts. However, these measures are unable to detect the kind of ontological relation or direction of the relation between them. For example, two concepts c and c' can be in hypernym-hyponym or domain-range kind of relation. The direction of relation can be (c, c') i.e. c is the hypernym and c' is the hyponym or (c', c) and so on. Further, the usage of contextual information of concepts in the statistical measures is also limited.

Hence a set of novel statistical scoring measures has been proposed to overcome these drawbacks. The proposed measures distinguish the kind and direction of relations using the set of keywords (Table I) and the relation label of the concept. Further, to guarantee that the domains of c and candidate wildcard filler of c are same, the scoring measure also uses the contextual information of c. All the scoring measures are based on the textual information obtained from the MEDLINE database. For each new pattern, a set of abstracts A from the MEDLINE database is retrieved using a search string. It consists of the candidate wildcard filler i.e. new candidate concept or relation and the old concept ( a concept in the ontology) in the new patterns which are concatenated using the "AND" operator. Depending on the type of the candidate wildcard filler, the set of abstracts is analyzed in different methods and correspondingly differ scoring measures are designed which are detailed below.

First, the candidate new hyponym (NHO) or new hypernym (NHY) is validated using the HH-CS (**H**yponym **H**ypernym-**C**oncept **S**coring) measure. The objective of this measure is to calibrate the Strength Of Association(SOA ) of hyponym-hypernym relation between the pair (NHO, OHY (old hypernym) in the new hyponym pattern or (NHY, OHO) in the new hypernym pattern. To assess the strength, the set of sentences S from the set of MEDLINE abstract A which matches the new pattern is retrieved. Further, the set of sentences which have old and new concepts of the new pattern along with any keyword of Table I is also retrieved. It is because; same information in the text can be conveyed in different wordings and syntactic structure. Hence searching for the textual information with the single keyword of the new pattern could be insufficient. Therefore, for each keyword of Table I, the SOA is computed and the maximum among it is considered as the final value. Further, to ensure that the discovered new concept (NHO or NHY) fit into the domain of old concept (OHO or OHY) and the given ontology, the contextual information (hypernym, hyponyms or siblings) of the old concept is also used for HH-CS computation. Only the hypernym of the old concept is used for the computation, since, the set of hyponyms and siblings of the old concept may be null, but hypernym of old concept always exist. Therefore, the SOA of hyponym-hypernym relation between the new concept and hypernym of old concept (i.e. HYPERNYM (old concept)) is also computed.

The description of the HH-CS measure is detailed below. The input to the measure is of two type: (i) new hyponym NHO and old hypernym OHY (OHY, NHO) if the candidate wildcard filler is a hyponym and (ii) the old hyponym OHO and new hypernym NHY (OHO, NHY) if the candidate wildcard filler is a hypernym. The output is the value of HH-CS indicating the SOA of hyponym-hypernym relation in the new pattern. For explanation, let us consider the input type to be (OHO, NHY) and the corresponding formal definition of HH-CS is given in (2). The measure consists of two components. First (HH-CS-$P_{1i}$ ()) and second (HH-CS-$P_{1j}$ ()) component gives the SOA computed based on the keyword set Key-Set1 and Key-Set2 respectively. The maximum value among the computed values based on the two keyword sets is chosen as the HH-CS value.

The description of the first component HH-CS-$P_{1i}$ () is as follows. The SOA is calculated by two sub-components. The first and second sub-component computes the SOA between (OHO, NHY) and (HYPERNYM(OHO), NHY) respectively as already mentioned above. The first sub-component is defined as the ratio of number of sentences in S containing OHO, $Key_i \in$ Key-Set1 and NHY (n(OHO, $Key_i$, NHY)) in the same order to the product of number of sentences in S containing OHO, $Key_i \in$ Key-Set1 (n(OHO, Key)) and $Key_i$

$\in$ Key-Set1 and NHY (n(Key$_i$, NHY)) in the same order respectively. Ensuring same order of concepts and keywords in the set of retrieved sentences implies correct direction of relation. This is the reason, for introducing two components for computing the value of HH-CS, since the order of concepts with Key-Set1 is hyponym and hypernym and it is vice versa in Key-Set2.

Similarly for computing the second sub-component, hypernym of OHO (HYPERNYM(OHO)) should be considered instead of OHO in the above ration. The product of these two sub-components is assigned as the value of HH-CS-P$_{1i}$ (OHO,NHY). Likewise, same description holds good for the second component HH-CS-P$_{1j}$ () except that the keyword set used is Key-Set2 (key$_j$ $\in$ Key-Set2).

Second, the candidate new domain concept (ND) or new range concept (NR) is validated using the DR-CS (**D**omain **R**ange-**C**oncept **S**coring) measure. The objective of this measure is to calibrate the SOA of non-taxonomical relation (NTR) between the pair (ND, OR (old range concept) in the new domain pattern or (OD, NR) in the new range pattern. Similar to the HH-CS measure, the set of sentence S is collected from the MEDLINE abstracts and the value of DR-CS is computed as shown in (5) for the input pair (OD, NR) with NTR.

Finally, the candidate relation (Rel) is validated using the RS (**R**elation **S**coring) measure. The objective of this measure is to calibrate the SOA of old concept C and C' through the new relation Rel in the relation new pattern. The Rel can indicate a taxonomical or non-taxonomical relation. The words in the matched sentences corresponding to the Rel position can be one among the set of keywords in Table I or a verb. The former indicates a taxonomical relation and the latter indicates the non-taxonomical relation. Similar to the HH-CS measure, the set S is collected from the MEDLINE abstracts and the value of RS is computed as shon in (6).

## 2.5 Positioning of the new concepts and relations

Finally, the scored new patterns are grouped into three categories such as 'hyponym and hypernym', 'domain and range' and relation patterns for selecting the validate patterns. The 'hyponym and hypernym' new patterns with scoring above the threshold $\beta_{HH}$ are chosen as validate patterns. Similarly 'domain and range' and relation patterns with scoring above the threshold $\beta_{DR}$ and $\beta_R$ are chosen respectively. The new concepts and relations in the chosen patterns are placed in the appropriate positions of the ontology provided, the new insertion does not create any redundancy or inconsistency issues. The redundancy is checked using a set of string similarity measures such as "string equality" [31] and "synonym similarity" [31]. Further, the inconsistency issue is ensured by pellet reasoner of Protege (http://protege.stanford.edu/). The set of newly placed concepts is used to form seed patterns for the next iteration. This process repeats until no new concepts or relations are discovered in the iteration.

## 3. Results and Discussions

For evaluating the proposed method, four biomedical resources such as MeSH (**Me**dical **S**ubject **H**eading) 2014, GENIA term and event ontology and biology ontology (http://www.tamps.cinvestav.mx/~arios/docs/datasets.zip) are used. The evaluation of the evolved ontology (EO) is processed using two set of metrics which are discussed in detail below.

$$\text{HH-CS (NHY, OHO)} = \text{Max (Max(HH-CS-P}_{1i}\text{(NHY, OHO)), Max(HH-CS-P}_{2j}\text{(NHY, OHO)))} \qquad (2)$$

where,

$$\text{HH-CS-P}_{1i}\text{(NHY, OHO)} = \frac{\text{n (OHO,key}_i\text{, NHY)}}{\text{n (OHO,key}_i\text{)} * \text{n (key}_i\text{, NHY)}} * \frac{\text{n (HYPERNYM(OHO),key}_i\text{, NHY)}}{\text{n (HYPERNYM(OHO),key}_i\text{)} * \text{n (key}_i\text{, NHY)}} \mid (1 <= i <= 3)\, \&\, key_i \in Key-Set1 \qquad (3)$$

$$\text{HH-CS-P}_{2j}\text{(NHY, OHO)} = \frac{\text{n (NHY,key}_j\text{, OHO)}}{\text{n (NHY,key}_j\text{)} * \text{n (key}_j\text{, OHO)}} * \frac{\text{n (NHY,key}_j\text{, HYPERNYM(OHO))}}{\text{n (NHY,key}_j\text{)} * \text{n (key}_j\text{, HYPERNYM(OHO))}} \mid 1 <= j <= 7 \qquad (4)$$

$$\text{DR-CS (OD, NTR, NR)} = \frac{\text{n (OD, NTR, NR)}}{\text{n (OD, NTR)} * \text{n (NTR, NR)}} * \frac{\text{n (HYPERNYM(OD), NTR, NR)}}{\text{n (HYPERNYM(OD), NTR)} * \text{n (NTR, NR)}} \qquad (5)$$

$$\text{RS (C, Rel, C')} = \frac{n(C, \mathrm{Re}l, C')}{n(C, C') * n(\mathrm{Re}l)} \qquad (6)$$

### 3.1 Precision, Recall and F-measure

Before evaluating the proposed method, the parameters of it such as k, $\beta_{HH}$, $\beta_{DR}$ and $\beta_R$ are to be determined. The values of k for CE measure for GENIA Term, GENIA Event, SBO, and MeSH are fixed to be 20, 20, 100 and 1000 respectively. The threshold value $\beta_{HH}$, which produced maximum effectiveness for the proposed ontology evolution method measured through F-measure (an entity can be hyponym or hypernym) is determined to be 0.0007. Similarly, the value of $\beta_{DR}$ is determined to be 0.006. The same value of $\beta_R$ for taxonomical and non-taxonomical relation patterns produce erroneous results since evidence for the former is more prominent. Hence the value of $\beta_R$ for taxonomical relations termed as $\beta_{R1}$ is identified individually based on the F-measure where it is computed considering only taxonomical relations as entities. Similarly, $\beta_{R2}$ for non-taxonomical relations is computed with F-measure where only non-taxonomical relations are considered as entities. The experimentally determined $\beta_{R1}$ and $\beta_{R1}$ values are 0.0005 and 0.04 respectively.

To evaluate the proposed method using the precision, recall, and F-measure metrics, first, a set of concepts to be evolved in the given biomedical resources are discovered using the CE measure. Each of the chosen concept c uses the proposed bootstrapping approach to evolve it. The newly discovered components of c are compared with the old components of c existing in the biomedical resources for computing the values of the evaluation metrics.

To depict the proficiency of the proposed ontology evolution method to discover new component, it is compared with the following 3 baseline systems. For all the baseline systems, the concepts to be evolved are randomly selected (RS). 20, 20, 100 and 1000 random terms are selected from GENIA Term, GENIA Event, SBO and MeSH ontologies respectively. (i) *LSP matching method [22] (u*ses the LSP of Table I) (ii) *Church's mutual information statistical measure (ST) [32, 33]* and (iii) *Combined approach (LSP + ST).* Further, to prove the proficiency of the individual sub-methods of proposed system such as concept selection (CS) based on CE measure and pattern based bootstrapping approach (BOOT), different versions of the proposed system are created as follows. (i) Version 1 (CS + LSP + ST) : Concepts that need to be evolved are identified using the proposed CE measure and top k concepts are chosen. Then the LSP + ST combined approach is used to evolve the chosen concepts. (ii) Version 2 (RS + BOOT) : The concepts to be evolved are chosen randomly. Then the proposed bootstrapping approach is used to evolve the chosen concepts. (iii) Version 3 (CS + BOOT): The proposed ontology evolution method.

The experimental results comparing the baseline and proposed methods are shown in Table 2. As evident from the results, the proposed method (CS+BOOT) outperforms all baselines and proposed method versions due to the following reasons. (i) The improved correctness (measured through precision) is achieved by the precisely designed scoring measures of bootstrapping approach such as HH-CS, DR-CS, and RS using the evidence from MEDLINE database and due to the usage of contextual information in the scoring measures. (ii) Better completeness (measured through recall) is obtained by exploring more information from the web pages. This is made possible by the new patterns created by the bootstrapping approach from the few seed patterns.

However, the effectiveness of the CS method seems same as the RS method based on these measures. Hence a different set of metrics termed as ontology quality metrics are used to depict the ability of the proposed method in discovering new components which is discussed in the following section.

## 3.2 Ontology quality metrics

Set of ontology quality metrics used are (i) "Number of concepts" in the ontology, (ii) "Average depth" of the concept in the ontology, (iii) "Maximum depth" of the ontology, (iv) "Inheritance Richness" (IR) [33] and (v) "Relation Richness " (RR) [33].

The three baseline systems and three proposed system versions are deployed to evolve the four biomedical resources mentioned above and compared using the quality metrics (Table 3-5). The increased number of concepts (i.e., lexical richness) in the Eos (Evolved ontologies) is better than that of the GOs (Golden ontologies), as shown by the "number of concepts". Similarly, the structural richness (i.e., the increased numbers of hypernyms and hyponyms) of the EO is proved by metrics such as the "maximum depth" and "IR". The increased richness is due to the following contributions. (i) The proposed CE measure is able to precisely select concepts that have a high possibility to be evolved with new concepts and relations. As shown in results, CS method based on CE measure outperforms the RS selection method and (ii) The proposed bootstrapping approach is designed to explore the vast knowledge available on the web by creating new patterns which aided in better coverage of the knowledge resources. The RR value of Biology GO is 0.24 and RR values of other three GOs are 0 since there is no non-taxonomical relation. The RR values of the EOs using the baseline and proposed methods are shown in Table V. The increased value of RR indicates that the proposed system is the effective one in discovering domain and range concept and non-taxonomical relations. Because LSP and ST measures don't possess any specific method or scoring measure to discover domain and range concept and non-taxonomical relations.

## 4. CONCLUSION

In this paper, the automatic evolution of biomedical ontologies is achieved using novel concept selection measure: CE (CS) and bootstrapping approach (BOOT). Specifically, the proposed CE measure is designed to precisely identify the concepts which have a high possibility to be evolved based on the number of neighbours and its depth. Then, a new lexical-syntactic pattern based bootstrapping approach with novel statistical measures such as HH-CS, DR-CS, and RS is used to identify the new component of the concepts using the knowledge from web pages and MEDLINE database. The large data resource of the web is explored using the bootstrapping approach (BOOT) and accurate new components are discovered from it using the new scoring measures. The experimental results based on precision, recall, F-measure proves that the proposed bootstrapping approach outperforms the existing baseline systems in precisely identify the components of the concept. Similarly, experimental results based on ontology quality metrics depict that the proposed ontology evolution method (CS + BOOT) discover more number of new components in comparison with the existing baseline systems. In future, a new methodology to predict the set of concepts or part of ontology which needs to be evolved in near future based on the trending topic in that domain is planned.

**Table 2. Precision and Recall of the proposed and baseline methods**

| | **Baseline** | | | | | | **Proposed Method** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **RS+LSP** | | **RS+ST** | | **RS+LSP+ST** | | **CS+LSP+ST** | | **RS+BOOT** | | **CS+BOOT** | |
| | P* | R* | P | R | P | R | P | R | P | R | P | R |
| GENIA Term | 0.48 | 0.52 | 0.44 | 0.56 | 0.62 | 0.5 | 0.63 | 0.47 | 0.74 | 0.70 | 0.75 | 0.69 |
| GENIA Event | 0.49 | 0.5 | 0.43 | 0.58 | 0.64 | 0.42 | 0.62 | 0.49 | 0.74 | 0.67 | 0.72 | 0.69 |
| Biology | 0.52 | 0.54 | 0.49 | 0.6 | 0.65 | 0.53 | 0.64 | 0.52 | 0.78 | 0.73 | 0.78 | 0.72 |
| MeSH | 0.47 | 0.45 | 0.42 | 0.47 | 0.57 | 0.4 | 0.59 | 0.42 | 0.68 | 0.58 | 0.69 | 0.59 |
| *P – Precision, R- Recall | | | | | | | | | | | | |

**Table III. Quality metrics of GOs**

| Ontology | NC | D | IR | RR |
|---|---|---|---|---|
| GENIA Term | 46 | 6 | 0.91 | - |
| GENIA Event | 36 | 6 | 0.92 | - |
| Biology | 172 | 5 | 1.33 | 0.24 |
| MeSH | 305349 | 15 | 1.42 | - |
| NC – Number of concepts, D – Depth, IR- Inheritance richness, RR – Relation richness | | | | |

**Table IV. Quality metrics of Eos**

| | Baseline | | | | | | | | | Proposed Method | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RS+LSP | | | RS+ST | | | RS+LSP+ST | | | CS+LSP+ST | | | RS+BOOT | | | CS+BOOT | | |
| | NC | D | IR | NC | D | IR | NC | D | IR | NC | D | IR | NC | D | IR | NC | D | IR |
| GENIA Term | 49 | 6 | 0.96 | 52 | 7 | 1.02 | 50 | 6 | 0.99 | 53 | 7 | 1.04 | 58 | 7 | 1.11 | 60 | 8 | 1.18 |
| GENIA Event | 38 | 6 | 0.97 | 41 | 7 | 1.02 | 39 | 6 | 1 | 42 | 6 | 1.06 | 46 | 7 | 1.15 | 49 | 7 | 1.18 |
| Biology | 184 | 6 | 1.43 | 195 | 6 | 1.51 | 186 | 5 | 1.43 | 198 | 5 | 1.52 | 215 | 6 | 1.62 | 222 | 6 | 1.7 |
| MeSH | 323349 | 16 | 1.5 | 344607 | 17 | 1.61 | 332749 | 16 | 1.54 | 348149 | 17 | 1.62 | 384739 | 18 | 1.76 | 393890 | 19 | 1.9 |

**Table V. RR value of EO of Biology ontology**

| Baseline | | | Proposed Method | | |
|---|---|---|---|---|---|
| RS+LSP | RS+ST | RS+LSP+ST | CS+LSP+ST | RS+BOOT | CS+BOOT |
| 0.25 | 0.27 | 0.26 | 0.28 | 0.29 | 0.31 |

## 5. REFERENCES

[1] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, J. Nucl. Acids. Res. 32, 267–270 (2004).

[2] BM. Konopka, Biomedical ontologies—A review, J. Biocybernetics and Biomedical Engineering 35, 75-86 (2015).

[3] The Gene Ontology Consortium, The Gene Ontology project, Nucl Acids Res 36, 440–444 (2008).

[4] AM. Khattak, R. Batool, Z. Pervez, AM. Khan and S. Lee, Ontology Evolution and Challenges, J. Inf. Sci. Eng 29, 851-71 (2013).

[5] AM. Khattak, K. Latif, S. Lee and YK. Lee. Ontology evolution: a survey and future challenges. Proceedings of the International Conference on U-and E-Service, Science and Technology, (2009) 68-75; Springer Berlin Heidelberg.

[6] G. Flouris, D. Manakanatas, H. Kondylakis, D. Plexousakis and G. Antoniou, Ontology change: Classification and survey, The Knowledge Engineering Review 23, 117-52 (2008).

[7] MC. Klein, Change management for distributed ontologies (2004).

[8] NF. Noy, A. Chugh, W. Liu and M.A. Musen. A framework for ontology evolution in collaborative environments. Proceeding of the International semantic web conference, (2006) 544-558; Springer Berlin Heidelberg.

[9] F. Zablith. Ontology evolution: a practical approach. Workshop on Matching and Meaning at Artificial Intelligence and Simulation of Behaviour (2009).

[10] V. Parekh, J. Gwo and T.W. Finin. Mining Domain Specific Texts and Glossaries to Evaluate and Enrich Domain Ontologies. Proceeding of the IKE (2004) 533-540.

[11] T.F. Gharib, N.L. Badr, S. Haridy and A. Abraham, Enriching Ontology Concepts Based on Texts from WWW and Corpus, J. UCS., 18, 2234-2251 (2012).

[12] B. Fortuna, M. Grobelnik and D. Mladenic D. Semi-automatic data-driven ontology construction system. Proceedings of the 9th International multi-conference Information Society IS-2006, Ljubljana, Slovenia (2006) Oct 9 223-226.

[13] G. Flouris G and D. Plexousakis, Handling ontology change: Survey and proposal for a future research direction, Institute of Computer Science, Forth. Greece, Technical Report TR-362 FORTH-ICS. (2005).

[14] P. Plessers P and O. De Troyer. Ontology change detection using a version log. Proceedings of the

International Semantic Web Conference 2005 Nov 6 (pp. 578-592). Springer Berlin Heidelberg.

[15] D. Rogozan D and G. Paquette. Managing ontology changes on the semantic web. Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (2005) 430-433. IEEE Computer Society.

[16] M. Klein, A. Kiryakov, D. Ognyanov, D. Fensel. Finding and characterizing changes in ontologies. Proceedings of the International Conference on Conceptual Modeling (2002) 79-89. Springer Berlin Heidelberg.

[17] S. Castano, A. Ferrara and S. Montanelli, Matching ontologies in open networked systems: Techniques and applications, J. on Data Semantics, 25-63 (2006).

[18] A.M. Khattak, K. Latif, S. Khan and N. Ahmed. Managing change history in web ontologies. Proceedings of the Fourth International Conference on Semantics, Knowledge and Grid (2008) 347-350.

[19] S. Castano, A. Ferrara and G.N. Hess. Discovery-Driven Ontology Evolution. Proceedings of the SWAP (2006).

[20] A. Kilgarriff and C. Fellbaum. WordNet: An Electronic Lexical Database (2000).

[21] S. Thenmalar, B. Sathiya B and T. V. Geetha, Learning concepts and relations for incremental ontology learning, Advances in Natural and Applied Sciences, 145-50 (2015).

[22] K. Liu, W.W. Chapman, G. Savova, C. G. Chute, N. Sioutos and R.S. Crowley, Effectiveness of lexico-syntactic pattern matching for ontology enrichment with clinical documents, Methods of information in medicine 50, 397- 407 (2011).

[23] Y. Zhu, M. Song and E. Yan, Identifying Liver Cancer and Its Relations with Diseases, Drugs, and Genes: A Literature-Based Approach, PloS one. 11, e015609 (2016).

[24] N. Collier, H.S. Park, N. Ogata, Y. Tateishi, C. Nobata, T. Ohta, T. Sekimizu, H. Imai. K. Ibushi K and J. I. Tsujii. The GENIA project: corpus-based knowledge acquisition and information extraction from genome research papers. Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics (1999) 271-272.

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

[26] H. W. Chun, Y. Tsuruoka, J. D. Kim, R. Shiba, N. Nagata, T. Hishiki and Jun'ichi Tsujii. Extraction of gene-disease relations from Medline using domain dictionaries and machine learning. Proceedings of the Pacific Symposium on Biocomputing (2006) 4-15.

[27] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. Proceedings of the 14th conference on Computational linguistics-Volume 2 (1992) 539-545. Association for Computational Linguistics.

[28] R. Snow, D. Jurafsky and A. Y. Ng, Learning syntactic patterns for automatic hypernym discovery, Advances in Neural Information Processing Systems 17, (2004).

[29] P. Cimiano, A. Hotho A and S. Staab, Learning concept hierarchies from text corpora using formal concept analysis, J. Artif. Intell. 24, 305-339 (2005).

[30] X. Jiang and A. H. Tan, CRCTOL: A semantic- based domain ontology learning system, J. of the American Society for Information Science and Technology 61, 150-68 (2010).

[31] J. Euzenat and P. Shvaiko, Ontology matching, Heidelberg: Springer (2007).

[32] K. Liu. Ontology Enrichment from Free-text Clinical Documents: A Comparison of Alternative Approaches (Doctoral dissertation, University of Pittsburgh).

[33] K. W. Church and P. Hanks, Word association norms, mutual information, and lexicography, J. Computational linguistics 16, 22-29 (1990).

[34] S. Tartir, I. B. Arpinar, M. Moore, A. P. Sheth and B. Aleman-Meza. OntoQA: Metric-based ontology quality analysis.