ORIGINAL ARTICLE

A Bootstrapping Approach for Entity Linking from Biomedical Literature

U. Kanimozhi, D. Manjula

Department of Computer Science and Engineering, College of Engineering, Anna University, Chennai, Tamil Nadu, India

Abstract

Aim: Entity linking (EL) is a task of aligning literal of a named-entity from an unstructured document to appropriate entities in a knowledge base. The main objective of EL in biomedical domain stems on the construction of efficient computational models. **Methodology:** The development corpus is a subset of PubMed and Medline abstracts dealing with Huntington disease and its genes. It was annotated with disease and gene relations, based on “etiology” and “clinical biomarker.” The input corpus consists of text related to Huntington disease, gene names with their functions and all words related to neurogenetic disorders. The input corpus which is manually curated has 8998 sentences and 140,481 words. **Results and Discussion:** A bootstrap approach based on uniformity perception and similarity computation to link entities from unstructured biomedical texts to ontologies. A rich semantic information and structures in ontologies are influenced by the proposed approach for similarity computation and entity ranking. **Conclusion:** The proposed approach addresses the EL in the biomedical domain. The experiments show that our approach outperforms the existing state-of-the-art algorithms in terms of linkage accuracy.

Key words: Biomedical literature, entity linking, bootstrapping

BACKGROUND

Over the past years, there is an emergence of enormous amount of unexplained abbreviations and terminologies that leverages a major bottleneck in understanding scientific literature. Mining and linking significant facts/information from biomedical literature have great impact on knowledge discovery in biomedical domain. It is also very challenging even for domain experts to keep up with the large number of articles published.[1] For instance, supporting the modeling task by means of identifying the key proteins, and their behaviors and interactions. Hence, there is a need for advancements in methodologies for making sense of a large amount of unstructured textual data which is explosively increasing. To facilitate it, a specific way of analysis can be enabled, where phrases comprising of a distinct term or sequence of terms are automatically linked to entries in a knowledge base (KB).

Here, the focus is on the task of entity linking (EL) from biomedical literature. EL is a process that links different entities that refer to the same source of data. Such entities exist in many other fields such as semantic web, multimedia, personal profiling, publication, and geography. Our main aim is to automatically identify the prominent entity mentions from unstructured texts and linking them to terms described in a KB and define in an ontology to enrich the text documents. These KB and ontology terms are also referred to as reference entities. EL can helps human end user navigate biomedical literature and improve many other natural language processing (NLP) tasks such as gene-disease association, gene-gene, and protein-protein interaction event extraction.[2,3] Entities enable semantic exploration of biomedical mentions; numerous information prerequisites can be encountered by recurring a list of entities, their properties, and/or their relations. Those entities can be utilized to identify unforeseen relations or functions and link the gap between unstructured and structured data.

Some recent works have been done on improving linkage performance using machine learning techniques.[4,5]

Address for correspondence:
U. Kanimozhi, Department of Computer Science and Engineering, College of Engineering, Anna University, Chennai, Tamil Nadu, India.
Phone: +91-8754285848/9940100060.
E-mail: kanimozhiu.03@gmail.com

Received: 02-02-2017
Revised: 13-02-2017
Accepted: 19-02-2017

However, it is laborious and effortful in building large-scale high-quality training set. Hence, we propose a bootstrapping approach for entity linkage by utilizing semantics-based and similarity-based methods. For a given entity, our approach initially infers a set of semantically co-referent entities and then, iteratively expands this entity set using distinct classes. To improve the performance of the classifier bootstrapping[6] technique is used which is suitable for entity linkage due to the abundant uncertain entities. A publicly accessible ontologies in a biomedical domain known as BioPortal[7] is utilized here. These ontologies consist of rich structures with declaratively defined semantic relations, along with comprehensive text descriptions provided by domain experts. We assume that multiple entities are semantically related in unstructured texts (i.e., they co-occur in the same sentence, are linked through dependency paths, or play certain semantic roles in the same event, etc.). Thereby address EL by means of uniformity perception by leveraging the global topical coherence and linking a set of relevant mentions simultaneously and generated labeled EL data through bootstrap approach.

In general, there are two categories of EL algorithms namely collective inference and non-collective methods, respectively.[1] Collective inference approaches influence concept mentions through supervised or graph based re-ranking methods. Besides they discourse the linking problem through exploiting the agreement between the mention document's text and the context of the entities of the KB. Graph-based re-ranking models typically collects linking agreement information from training data and propagates to other nodes. Non-collective methods usually rely on prior knowledge and context similarity with supervised models. Ranking scores for each concept mentions are computed individually. Whereas, both these approaches requires large amount of manually labeled entity mentions to achieve a reasonable linking accuracy.

This paper presents a study on identifying prominent links between entities and label gene/protein-disease relations in PubMed and Medline abstracts by using bootstrapping approach. Beginning with PubMed and Medline abstracts, we first recognized gene/protein and disease entities using existing NLP tools such as Regex-NER. Then, we extract candidate gene/protein-disease pairs by mapping it with the existing ontology and KB based on different levels of co-occurrence such as abstract level, sentence level, phrase level, and paragraph level. To find the most linked gene/protein-disease relations, we finally rank candidate gene/protein-disease pairs using information gain (IG). The evaluation using a manually annotated dataset from gene ontology (GO) indicated that the bootstrap method outperformed others. To the best of our knowledge, this is the first attempt that applied bootstrap approach to rank gene/protein-disease EL from biomedical literature.

METHODOLOGY

Proposed framework

A bootstrapping approach is proposed, in which it accepts the candidate entity as an input. The following section describes the methodological steps for our approach depicted in Figure 1. The major goal of the paper is to link the identified entities to the concepts in the KB. For a given biomedical text document as input, we extract the entity mentions and automatically construct a kernel, consisting of semantically co-referent entities, through uniformity perception on several gene/protein/disease vocabularies from ontologies. Then the kernel is expanded iteratively using distinct classes to probe different co-referent entities. The distinguishability of each class is learned with a statistical measure, revealing the importance of the class characterizing the co-referent entities and match the class by comparing the functions of those entities. Furthermore, frequent class combinations (i.e., the functions often used together) are mined to enhance entity labeling criterion in bootstrapping, so that the linkage accuracy can be improved. Entity mapping for an entity mention is assigned by measuring the popularity of an entity among all other candidate entities.

Input documents

The development corpus is a subset of PubMed and Medline abstracts dealing with Huntington disease and its genes. It was annotated with disease and gene relations, based on "etiology" and "clinical biomarker." Beginning with PubMed and Medline abstract collection. The initial step is the preprocessing which is done to determine entity boundaries in a text by sentence splitting and tokenization. NLP incurs creation of a set of patterns to match the possible linguistic realizations of the individual facts. Due to this complexity, the preprocessing on structural input requires assigning parts-of-speech and features to words and idiomatic phrases. Annotated corpus drive construction of training data for machine learning that will filter out false positives from the dictionary-based results. These data are used for training and testing purposes. The input corpus consists of text related to Huntington disease, gene names with their functions and all words related to neurogenetic disorders. The input corpus which is manually curated has 8998 sentences and 140481 words. Let, entity mentions $u \in E$ are prominent phrases in the input biomedical text document. All classes, properties and individuals described in the ontologies $r \in R$ are considered to be the reference entities. Relations based on sentence level and paragraph level are extracted based on co-occurrence are extracted. A list of candidate entities X is located from the biomedical dictionaries for each entity in the context graph GG. Then we compute the importance score and link them by the bootstrap approach. Finally, we compute similarity scores for each entity/candidate pairs $<u, x>$ and select the candidate with the highest score as the appropriate entity for linking.

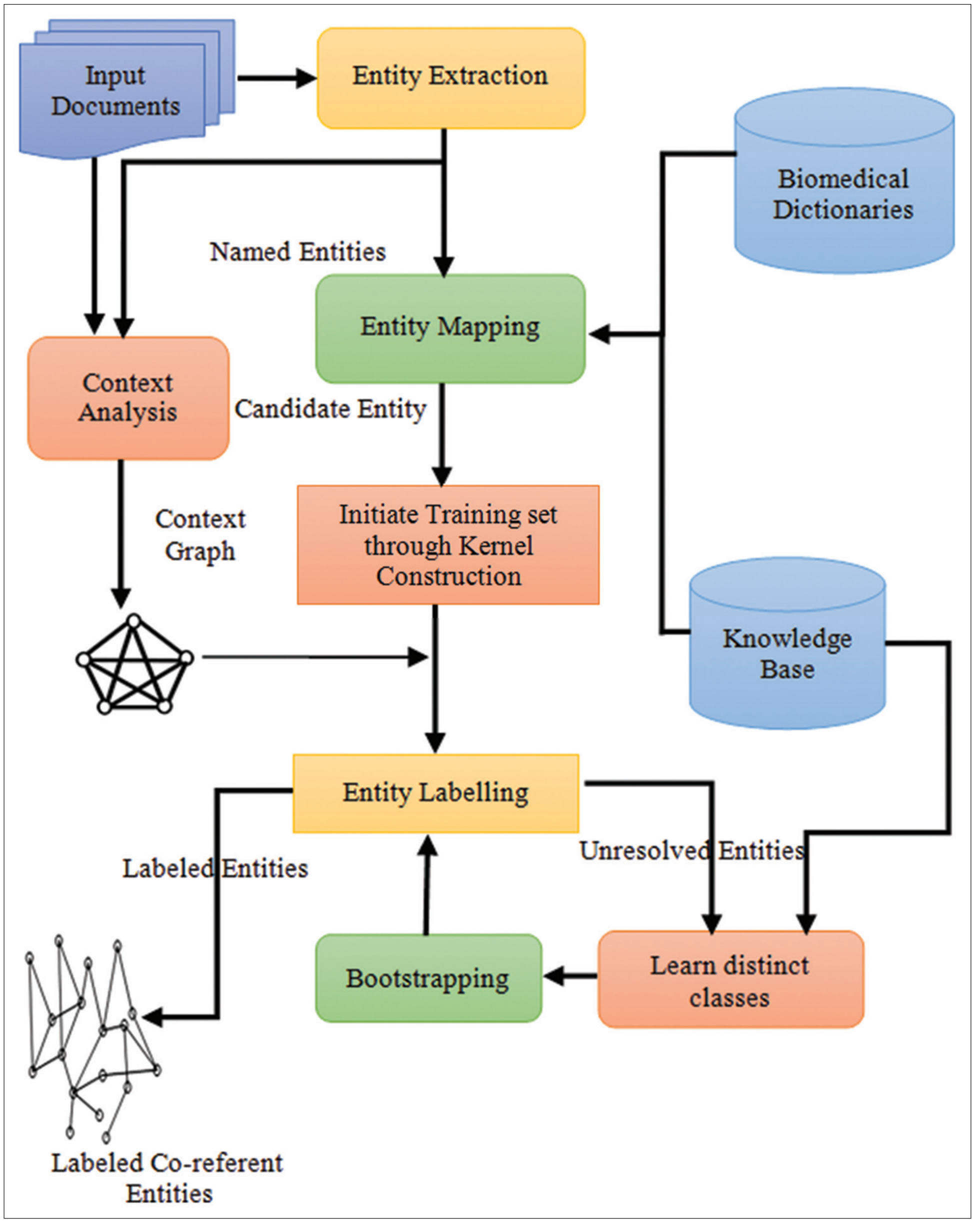

Figure 1: Work flow of the bootstrap entity linkage

Entity extraction

We apply the publicly available NLP tools for identifying prominent biomedical entities from unstructured texts to recognize the entity and ascribe it to a class or entity type. The occurrences of gene/disease entities in a text automatically identified by gene/disease NER. Initially, a name tagger[8] is used to extract entity mentions. Regular expressions are used to join named entities that might have been considered separate by looking for intervening prepositions, articles, and punctuation marks. After that, a shallow parser[9] is used to add noun phrase chunks to the list of entities. A parameter controls the minimum and maximum number of chunks per entity, in which by default one and four are considered and whether overlapping entities are allowed. The entity normalization process is characterized by representing entities' names to their canonical names and by associating them with unique representations so as to help in solving issues resulting from variations in the synonym terms as well as the ambiguous abbreviations.

KB

A comprehensive KB is developed based on the classes, characters and functions/properties present in the aggregated ontologies. Graph-based approach is used to construct the KB. We create a document for triplet construction in which each entity is entity is described as a set of triplest ∈ T. The KB which is constructed from 300 biomedical ontologies from BioPortal,[7] consist of the Triplets in which each entity is connected to other entities via a set of triples T. Moreover, these connections are regarded as edges of CG_{KB} where, CG_{KB} is the context graph with respect to the KB.

Entity mapping and candidate retrieval

We perform entity matching for all entity types based on regular expressions[10] using Regex-NER. It defines cascaded patterns over token sequences. Set of rules are defined for each entity type that expresses some patterns of entity mentions by exploring the corpora, and BIO labels are assigned to those patterns. Also, triples describing the entities are analyzed based on the properties such as: Labels and names (e.g., rdfs: label), synonyms (e.g., synonym from GO), aliases, and symbols (e.g., from orphanet ontology). Thus, providing more than 160 properties to map with its respective entity. Then, we retrieve all the entities that are similar to the mentions in the ontologies and KB and consider them as candidate entities.

Kernel construction

A set of semantically co-referent entities u mentioned as kernel of u is automatically inferred by using the functional aspects of gene and disease mined from biomedical dictionaries. We use human metabolome database,[11] GO[12] and UniProt[13] as gene dictionary; medical subject heading produced by the US National Librart of Medicine, and Kyoto Encyclopedia of Genes and Genomes disease[14] as disease dictionary. The training set is initialized by combining the candidate entity with the co-referent entities based on functional property, partial match and full match of the elements. Besides, we assume that the correct entities are infers in the kernel. Yet, error accumulation in the bootstrapping process can be encountered due noisy data.

Entity labeling

Linking entities refers to the description of functions or target genes through which it is associated with the disease. Entity labeling is a task that deliberated for context graph in our experiments. Assume that the classification component in a given context graph is given a set of labels. The problem is simplified by initially constructing a network with only a single type of entity from the context graph. So that we introduce a link between two entities if they are connected to the same function and having introduced these entity-entity links delete all the functional nodes and the links originating from them from the context graph.

Learning distinct classes

This iterative step is based on the hypothesis that co-referent entities share some similar functional aspects and a few functions are more essential for linking entities. For a given set of candidate entities with respect to u, we estimated a set of co-referent and non-coreferent entities together establish the training set of u. A pair of matched functions (partial/full) are chosen to hold the maximum distinguishability and is measured in terms of IG.[15] Then assigned a unique value to separate function in that class. Since functional relations are involved in the iteration. Functional relations are extracted and compared with a string matching algorithm[16] for the entities given in the training set. Since each entity is described in the dictionary that contains all phrases matching the string. If the similarity between the values is larger than a threshold, the related two functions are matched. The highest computational cost in the boosting process is incurred due to functional aspect comparison. The learned functional classes reveal important characteristics of the mind biomedical literature and enable to find new co-referent entities holding the same function. In addition, we employ Apriori algorithm to find the frequent grouping of functions and refine them using heuristic rules beforehand. In each iteration, when a class is chosen, and it belongs to some frequent class cluster, its counterpart in the group. Finally, the classes in the group with their associated value would be used together to obtain new links.

Bootstrapping algorithm

The proposed entity linkage algorithm is a kind of semi-supervised learning and is depicted in Algorithm 1. Given the kernel K(e) of an entity e, and a set X of (uncertain) entities from the input document D, the goal is to incrementally learn the most distinct classes (Steps 3-5) and use them to continue linking entities in X by retraining itself on an explained training set (Steps 7-8).

Algorithm 1: Bootstrapping biomedical entity linkage

```
Input: The kernel K of an entity e, a set X of candidate entities in a set D of input documents.
Output: A set E of labeled co-referent entities fore.
    1. Initialize two empty lists $L_p$ and $L_v$, and print K to E
    2. Estimate a set N of non-co referent entities for e
    Such that $N \subseteq X, |N| \approx |E|$;
    3. The most distinct classes are selected $(f_i, f_j) \notin LP$ by
    District $(f_i, f_j) = IG(f_i, f_j) = E(T) - E(T_{(fi, fj)})$, such that
    $f_i \in U_{s \in EUN} Pred(D, s), f_j \in U_{t \in EUN, t \neq s} Pred(D,t)$;
    if $(f_i, f_j)$=NULL then break;
    4. Assign the maximum score values $(v_i, v_j)$ to $(f_i, f_j)$ respectively, based on occurrence given by
    $(v_i, v_j)=\text{argmax}|\{(s, s') \in E \times E | sub(v, v') \geq \delta, \langle s, f_i, v \rangle \in D, \langle s', f_j, v' \rangle \in D\}|$, such that
    5. $v_i \in \bigcup_{s \in E} Obj(D,s,f_i)$, $v_j \in \bigcup_{t \in E} ObjD,t,f_j$, while $(F_i,v_i) \notin L_V$ or $(f_j,v_j) \notin L_{V;}$
    6. If $o_i$=NULL and $o_j$=NULL then Push $(f_i,f_j)$ in $L_{P;}$ Go to step 3;
    7. If $o_i \neq$NULL then
Use $(f_i, v_i)$ to fetch out a set U of candidate entities, such that
$U \leftarrow \{u \in X \mid \langle u, f_i, v_i \rangle \in D\}$;
Else if $(f_i, v_i)$ is distinct by the following equation:
$\text{Distinct}\left(f_i, v_i\right) = \frac{|\{s \in \mathbf{E} | s, f_i, v_i \in \mathbf{D}\}|}{|\{s \in \mathbf{X} | s, f_i, v_i \in \mathbf{D}\}|}$, then
Add U to E, and eliminate U from X;
```

```
    8. If $o_j \neq NULL$ then
    Use $(f_j, v_i)$ to draw a set W of candidate entities, such that
    $W \leftarrow \{u \in X \mid \langle w, f_j, v_j \rangle \in D\}$;
    Else if $(f_j, v_i)$ is distinct by the following equation,
$\text{Distinct}(f_i, v_i) = \frac{|\{s \in \mathbf{E} \mid s, f_i, v_i \in \mathbf{D}\}|}{|\{s \in \mathbf{X} \mid s, f_i, v_i \in \mathbf{D}\}|}$, then
Add W to E, and eliminate W from X;
    9. Push $(F_i, v_i)$ and $(F_i, v_i)$ in $L_v$;
    10. Continue iteration until iteration_times>$\tau$; finally,
        return E with the set of labels.
```

The distinctiveness of a class is measured with relation to the amount of potentially co-referent entities that can be found using the class. Let (f_j, v_i) be the distinct function and value selected from a set D of documents, respectively. The distinctiveness of a class is computed as mentioned in Steps 7 and 8, where E, X are the co-referent and uncertain entity sets in D, respectively. The algorithm terminates when all distinct functions have been checked (Step 7) or the iteration time exceeds the threshold value τ (Step 10). A subset of E is randomly sampled by the algorithm to reduce the computational cost. The sample size is denoted by N is set to set to 240 based on the computational capability of our system. The time complexity of the algorithm $O(\tau^*N)^2$, since in an iteration at most O(N) entities need to be compared, which is the most time-consuming step in the algorithm and the main issue of our approach.

EXPERIMENTAL RESULTS

Evaluation measures

The assessment of our EL systems is performed in terms of evaluation measures, such as precision, recall, F_1-measure, and accuracy. The precision of an EL system is computed as the portion of correctly linked entity mentions that are generated by the system:

$$\text{Precision} = \frac{|\{\text{Correctly linked entity mentions}\}|}{|\{\text{Correctly linked entity mentions}\}|} \quad (1)$$

Precision takes into account all entity mentions that are linked by the system and determines how correct entity mentions linked by the EL system are. Precision is usually used with the measure recall, the portion of correctly linked entity mentions that should be linked:

$$\text{Recall} = \frac{|\{\text{Correctly linked entity mentions}\}|}{|\{\text{Entity mentions that should be linked}\}|} \quad (2)$$

Recall takes into account all entity mentions that should be linked and determines how correct linked entity mentions are with regard to total entity mentions that should be linked. These two measures are used together in F_1-measure to provide a single measurement for a system. F_1-measure is defined as the harmonic mean of precision and recall:

$$F_1 = \frac{2.\text{precision.recall}}{\text{precision+recall}} \quad (3)$$

Accuracy is calculated as the number of correctly linked entity mentions divided by the total number of all entity mentions. Therefore, here precision=recall=F_1=accuracy.

RESULTS

The experimental evaluation was performed on a personal desktop with an Intel Core i5 3.1 GHz CPU, 4 GB memory, Ubundu 11.10 and Java 7. The datasets were stored on a server with two Xeon Quad 2.4 GHz CPUs, 64 GB memory, CentOS 6.4 and MySQL 5.6. We conducted our experiment using the evaluation dataset created by Zheng *et al.* (2015) which contains 208 linkable mentions extracted from several biomedical publications. Among all of the ontologies, there are more than 2 million entities and more than 50 million factual statements. We observed that for each mention, the candidate entity types are not as diverse. The kernel achieved the highest precision but the lowest relative recall because some co-referent entities cannot simply be identified through uniformity perception. During bootstrapping, our approach estimated non-coreferent entities to measure distinctiveness and employed a frequent combination of functions/relations between entities to enhance the selection criterion of functions/relations. The candidate "neural nucleus" a non-coreferent entity indirectly links to "nerve impulse," due to frequent combination of relations it links to "neural nucleus" from candidates of co-referent entities enables the candidate entity "cell nucleus" to obtain the correct label and rank to link. Both of their contributions increased the overall accuracy of the proposed system. It is observed that 64% of the correct links were inferred from the kernel, and 36% correct links were established through bootstrapping, in which 4% were of frequent combinations of relations.

The EL is to map an entity mentioned in an input text to the KB, which consist of articles from PubMed and Medline. The Bootstrap track gives a sample entity set which consists of 416 entities for developing. The test set consists of 3904 entities. 2229 of these entities cannot be mapped to KB, for which the systems should return NIL links. The remaining 1675 entities all can be aligned to KB. We will first analyze the ranking methods with those non-NIL entities, and then with an additional validation module, we train and test with all entities including NIL link entities. Table 1 provides the comparative analysis with the existing EL systems. We have shown our results for Biomedical EL system before and after bootstrapping of the biomedical entities in Table 2.

For example, given a sentence "The effects of the MEK inhibitor on total HER2, HER3 and on phosphorylated pHER3 were dose dependent." it can link "HER2" to "ERBB2" in BioPortal and extract the class.

"Proto-Oncogenes→Oncogenes→Genes→GenomeComponents→Genome→Phenomena and Processes" as the class and label the co-referent entities for this entity mention.

Figure 2 depicts the average precision and relative recall on the 50 testing entities with respect to the number of iterations, where the relative recall continuously rises up at the beginning and ascends slowly later. The result suggests that a small amount of distinct functions is accurate enough for entity linkage. If bootstrapping continues, some non-distinct functions would be chosen and cause a decrease in precision. Based on Figure 2, we set the maximum number of iteration $\tau = 4$.

The empirical comparison has been made between the proposed Bootstrap approach and two other systems, which have several variations and it is hard to cover them all in our test. The indexing + similarity computation approach[17] leverages indexing techniques on a few important relations/functions to locate the candidate entities and then combines various matchers to compute similarities between these candidates. In our evaluation, we indexed the labels and mention names of all the entities in our input corpus and used the TF_IDF model to compute the similarities between the descriptions of entities. The similarity threshold was set to 0.24 based on the best accuracy in our test. Class-based learning approach identifies distinct functions/relations statistically with relation to different classes, and matches other entities under the same classes using the learned function/relations. We chose[5] in the test, which conducted uniformity perception to create a training set and ranked entities with relation to the IG in different classes. Value similarities from top-5 function/relations were linearly aggregated with equal weighting, and the threshold was fixed to 0.14 according to the best accuracy.

Figure 3 shows the precision and relative recall comparison between the proposed Bootstrap approach and the other two systems. It is observed that the Bootstrap approach achieved the best overall accuracy, while the class-based learning largely depend on the sufficiency of the training sets, causing its accuracy varied between testing entities. The system based on indexing + similarity computation performed worse than the others, because it generated too many candidate entities with diverse functions/relation-values, and failed to decide a uniform threshold to eliminate wrong links.

Table 1: Performance of the entity linking systems

EL system	Correct links	Total links	Linkage accuracy (%)
Chan and Roth (2013)	84	113	74.34
Zheng *et al*., (2015)	173	208	83.17
Bootstrap approach	192	208	92.30

EL: Entity linking

Table 2: Results for bootstrapping biomedical entity linking system

Method	Precision (%)	Recall (%)	F-score (%)
Without bootstrap	86.44	87.23	86.83
With bootstrap	92.83	83.12	92.19

EL: Entity linking

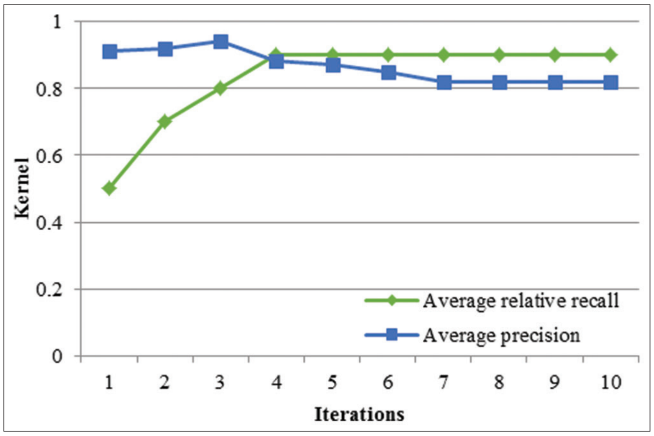

Figure 2: Precision and relative recall with relation to number of iteration

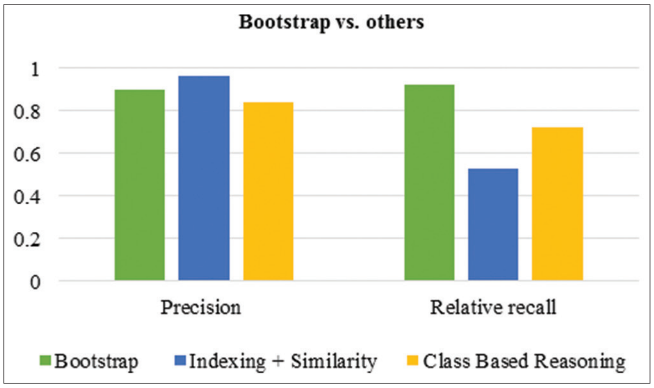

Figure 3: Precision and relative recall comparison

CONCLUSION

We proposed a bootstrapping approach to entity linkage on biomedical domain. It automatically extracts and link prominent entities from unstructured biomedical literature to ontologies. The proposed Bootstrap approach is based on uniformity perception and similarity computation to link entities from unstructured biomedical texts to ontologies. Furthermore, bridges the gap between semantically co-referent entities and potential candidates. The experimental results show that our approach achieved superior precision and recall by comparing with the existing state-of-the-art algorithms with improved linkage accuracy. In future, we look forward to designing other semi-supervised learning approaches for entity linkage over the biomedical domain.

REFERENCES

1. Hunter L, Cohen KB. Biomedical language processing: Perspective what's beyond PubMed? Mol Cell 2006;21:589-94.
2. Miwa M, Sætre R, Miyao Y, Tsujii J. A Rich Feature Vector for Proteinprotein Interaction Extraction from Multiple Corpora. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics; 2009. p. 121-30.
3. Liu B, Qian L, Wang H, Zhou G. Dependency-Driven Feature-Based Learning for Extracting Protein-Protein Interactions from Biomedical Text. Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Beijing, China: Association for Computational Linguistics; 2010. p. 757-65.
4. Isele R, Bizer C. Active learning of expressive linkage rules using genetic programming. J Web Semant 2013;23:2-15.
5. Hu W, Yang R, Qu Y. Automatically generating data linkages using class-based discriminative properties. Data Knowl Eng 2014;91:34-51.
6. Abney S. Bootstrapping. In: Proceeding Annual Meeting on Association for Computational Linguistics. East Stroudsburg, PA: ACL; 2002. p. 360-7.
7. National Center for Biomedical Ontology: Bio Portal; 2014.
8. Ratinov L, Roth D. Design Challenges and Misconceptions in Named Entity Recognition. Proceedings of the Thirteenth Conference on Computational Natural Language Learning Association for Computational Linguistics, Boulder, CO; 2009. p. 147-55.
9. Punyakanok V, Roth D. The use of classifiers in sequential inference. NIPS. Vancouver. British Columbia, Canada: University of Illinois at Urbana-Champaign; 2001.
10. Chang AX, Manning CD. Tokens Regex: Defining Cascaded Regular Expressions Over Tokens. Technical Report CSTR 2014-02, Department of Computer Science, Stanford University; 2014.
11. Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y, *et al.* HMDB 3.0 - The human metabolome database in 2013. Nucleic Acids Res 2013;41:D801-7.
12. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, *et al.* Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25-9.
13. UniProt Consortium. The universal protein resource (UniProt). Nucleic Acids Res 2008;36:D190-5.
14. National Library of Medicine. MeSH. Available from: http://www.ncbi.nlm. nih.gov/mesh.
15. Mitchell T. Machine Learning. New York: McGraw Hill; 1997.
16. Stoilos G, Stamou G, Kollias S. A String Metric for Ontology Alignment. Proceeding International Semantic Web Conference, ISWC'05; 2005. p. 623-37.
17. Rong S, Niu X, Xiang E, Wang H, Yang Q, Yu Y. A Machine Learning Approach for Instance Matching Based on Similarity Metrics. Proceeding International Semantic Web Conference, ISWC'12; 2012. p. 460-75.
18. Hu W, Yang R, Qu Y. Automatically generating data linkages using class-based discriminative properties. Data Knowl Eng 2014;91: 34-51.
19. Zheng JG, Howsmon D, Zhang B, Hahn J, McGuinness D, Hendler J, *et al*. Entity linking for biomedical literature. BMC Med Inf Decis Making 2015;15 Suppl 1:S4.

Source of Support: Nil. **Conflict of Interest:** None declared.