Constructing Dynamic Ontologies from Biomedical Publications

Megha Nagabhushan
School of Computing and Engineering
University of Missouri - Kansas City
Kansas City, Missouri 64110
Email: mnbpc@mail.umkc.edu

Rohithkumar Nagulapati
& Mayanka Chandrashekar
School of Computing and Engineering
University of Missouri - Kansas City
Kansas City, Missouri 64110
Email: rnd95, mckw9@mail.umkc.edu

Yugyung Lee
School of Computing and Engineering
University of Missouri - Kansas City
Kansas City, Missouri 64110
Email: leeyu@umkc.edu

***Abstract*—In recent years, there has been explosive growth in the amount of biomedical publications. In this paper, we propose a semantic framework that aims to automatically generate an ontology by extracting assertions and topics from multiple free-text scientific publications in PubMed. The pipeline approach for knowledge discovery and ontology generation in the proposed framework has been implemented on the Spark parallel engine based on the Stanford CoreNLP for Natural Language Processing, TF-IDF (Term Frequency Inverse Document Frequency) for feature extraction, OpenIE (Open Information Extraction) for relation extraction, K-Means clustering for topic discovery and OWL API for ontology generation. We have shown that the ontology generated may be very effective in biomedical applications (such as paper search and summarization) with scientific publications.**

I. Introduction

In recent years, there has been explosive growth in the amount of biomedical data being generated, with the majority being unstructured data. These publications could be extracted for novel findings and hypotheses from research (called 'assertions' in this paper). The dissemination and sharing of biomedical findings to translational medicine are not easy, even if some of the assertions are freely available through publications. The most up-to-date findings about diagnoses, interventions, and treatments would be important when attempting to make critical decisions for patients by physicians and researchers in health care.

There is an increasing demand for ontologies that can be used to expedite the discovery of new diagnostic treatments and interventions for medical applications such as knowledge retrieval, summarization, and medical question answering systems. Dynamic and relevant ontologies will be more useful for evidence-based medicine or personalized treatment than general ontologies.

Many existing ontologies are typically designed by domain experts. However, in Text2Onto [1], OntoLearn [15], and Sprat [9], semi-automatic methods were proposed for ontology construction from textual data. Unlike these ontologies, in this paper, we propose a semantic framework that aims to automatically generate an ontology by extracting assertions (e.g., facts, findings) from multiple free-text sources such as PubMed publications.

In order to extract ontological assertions from a free-text corpus, in [6], the typical steps are defined as follows: first, a subject or object will be recognized from a sentence using named-entity recognition (NER), PoS tagging or Information Extraction (OpenIE) techniques; second, a relationship between subjects and objects will be detected, and third, relations can be classified (e.g., generalization or specialization).

Unlike existing approaches, we design a semantic framework for converting free-text input (e.g., PubMed paper abstracts) to an ontology with assertions (e.g., subject, predicate, object). This was designed through a pipeline approach of (i) Natural Language Processing (NLP) [8] and Information Retrieval (TF-IDF) [5], (ii) Information Extraction (OpenIE) [14], (iii) Topic Discovery using Clustering [3], and (iv) Ontology Generation using OWL API [4].

II. Related Work

Many works have been proposed to discovery entities and their relations from medical corpus. Omura et al. [12] proposed to find the disease similarity by constructing a graph representing anatomical features using local information without considering any features unique to a disease. Chan et al. [7] classified the diseases by fetching the relationships between three categories of diseases. Our work is different from these works in terms of discovering the topics and their relationships through extracting common or unique assertions $<$ Subject, Predicate, Object$>$ from multiple medical domains (e.g., obesity, diabetes, heart diseases).

Lossio-Ventura [6] focused on an automatic construction of a knowledge base from heterogeneous information sources on Obesity. Name Entity Recognition (NER) and Relation Extraction (RE) were used to construct an ontology for a text corpus. In this work, they used the biomedical entity detection, which is very useful to determine the meaning of the various medical terms. A binary classification is used for extracting such relationships. Binary classification is not suitable in discovery of the multiple entities and their complex relations from a large corpus. In addition, they rely on domain experts' manual annotation of medical terms and relations. However, for a large text corpus, manual annotation would not be feasible. To overcome this limitation, in our work, K-Means

clustering [3] is to find clusters of related triplets <Subject, Object, Predicate> from the text corpus.

FRED [2] automatically generated RDF/OWL formatted ontologies from the texts. In FRED, each frame is to formalize verbs or other linguistic constructions as OWL n-ary relations according to Frame semantics. FRED supports representing modality, tense, and negation of sentences. However, it is incapable of handling large datasets for generation of ontology. Its results are not uniform if both facts and concepts are expressed by natural language text.

III. Proposed Framework

There are three primary components in the Knowledge Discovery Framework: (i) **Assertion Discovery**: The assertion is defined as facts or findings from biomedical research that will be defined as a form of <Subject(S), Predicate(P), Object(S)>; (ii) **Topic Discovery**: A topic is defined as underlying contexts of a given corpus that will be defined as a set of related assertions $< S_i, P_i, O_i¿, \ldots, < S_j, P_j, O_j¿$; (iii) **Ontology Generation**: A set of the related assertions (called topics) discovered from the previous components is represented in a form of ontology (OWL). Each component is described below.

A. Assertion Discovery

The assertion discovery is based on the integration process of the following two steps: (i) Natural Language Processing (NLP) citeCoreNLP and Information Retrieval (TF-IDF) [5]; (ii) Triplet Discovery using OpenIE [14].

1) Natural Language Processing (NLP): This step is conducted using CoreNLP Library [8]. The NLP tasks were conducted as follows:

- Tokenization: Tokenization is the process of breaking sentences into tokens that are the smallest constructs of any text data.
- Lemmatization: Lemmatization is the process of separating words into individual morphemes and identifying the class of the morphemes.
- Stopword Removal: Stopword Removal is the process of removing stopwords from the corpus. For example, the stopwords in English include able, about, above, according, accordingly, etc.

2) Feature Discovery using TF-IDF: The term frequency-inverse document frequency (TF-IDF) is a numerical statistic that is intended to reflect how important a word is to a document in a corpus. Using TF-IDF, the representative features were extracted using TF-IDF, which is the product of term frequency and inverse document frequency. The term frequency is the number of annotated terms that appear in a specific document (Equation 1). Document frequency is the frequency of the terms in all the documents. The Inverse Document Frequency intends to reduce the importance of the word that occurs most frequently in all the documents (Equation 2). It is mainly used to eliminate the common terms across all the documents. The IDF value is computed by dividing the total number of documents with the number of documents that contain the given term t and then by applying a logarithm to the resultant value. If the term appears in more than one corpus, it is more likely to be a common term that is not specific to any given document, and hence, the log value of the word is reduced to zero ensuring that the IDF value and thereby, the TF-IDF values are less for this term (Equation 3).

$$TF(t,d) = 1 + \log(f_{t,d}) \tag{1}$$

$$IDF(t,D) = \log \frac{N}{1 + |\{d \in D : t \in d\}|} \tag{2}$$

where N is the total number of the documents in the corpus, i.e., $N = |D|$ and $|\{d \in D : t \in d\}|$ is the number of documents where the term t appears (i.e., $TF(t,d) \neq 0$).

The TF-IDF value is high if the term has a high term frequency and a low document frequency in the corpus. Hence, by considering the TF-IDF value, we can eliminate the common terms.

$$TF - IDF(t,D) = TF(t,d) \cdot IDF(t,D) \tag{3}$$

The top X important words can extracted based on the weights of the TF-IDF terms (e.g., X= 50).

3) Triplet Discovery: Open Information Extraction (OpenIE) was performed on our corpus. Open information extraction (OpenIE) [14] extracted relation tuples, typically binary relations, from plain text and produced its result in the form of a Quadruple <Subject,Predicate,Object,Confidence-Score> from the triplet <Subject(S),Predicate(P),Object(O)>, where the Confidence-Score is higher than a threshold δ.

4) Assertion Discovery: We discover the important triplets <S,P,O> by matching the TF-IDF results with the OpenIE triplets whose terms match one of the top TF-IDF words. The Assertions <S,P,O> will be discovered if any given triplets from the previous step, the assertion <S,P,O> contains any top TF-IDF terms t (i.e., $t \in SPO$ where SPO is a concatenated word of <S,P,O> and t is a top TF-IDF).

B. Topic Discovery

Assertion Clustering: The first step in *Topic Discovery* is to cluster the assertions <SPO> from the Assertion Discovery component into groups of similar triplets. The second step is to find a topic name of the the topic cluster using a topic discovery technique. Due to space limitation, the discussion about the second step will be omitted from this paper. For clustering, we used the K-Means (KM) clustering algorithm [3]. In this paper, we revised the KM algorithm for discovery of relevant assertions <SPO> from publication abstracts to discover topics. The KM algorithm is an unsupervised learning technique for partitioning assertions <SPO> (subject, predicate, object) into *K* different clusters by clustering them with the nearest mean.

Given a set of the SPO terms ($SPO_1, SPO_2, \ldots, SPO_n$), where each paper can be represented as an SPO vector, KM clustering aims to partition the n terms into k, where $k \leq n$ and clusters $C = C_1, C_2, \ldots, C_k$, for minimizing the within-cluster sum of squares (sum of distance functions of each point in the cluster to the k center) (Equation 4).

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \mu_i\|^2 \quad (4)$$

where μ_i is the mean of points in S_i. We conducted a string matching operation with each triplet <SPO> against all other triplets to determine the similar triplets. For our experiment, we considered the size of clusters, $k = 10$.

C. Ontology Generation

In the ontology generation, the assertions and topics discovered from the previous steps need to be converted to a ontology in RDF (resource description framework) and OWL (ontology web language). For the formal representation, we need to represent them as schema and data.

1) Medical Term Discovery: This step is to match Subject (S) or Object (O) in <S, P, O> with existing medical ontologies using the NCBO BioPortal Annotator [11]. The ontologies in NCBO BioPortal include SNOMED, ONTOAD, MESH, and others. If S or O in <S, P, O> contain any medical terms, then these terms will be identified with their ontologies.

2) Ontology Learning: For our purposes, we designed the three steps as follows: Concept Learning, Property Learning, and Triplet Learning.

Concept Learning: The classes and their individuals will be identified from the Medical Term Discovery. A medical term from the Annotator and topic terms from the Topic Discovery are considered as a class. Individuals for these classes are also generated. Regarding the Concept Generation, we have two cases as follows: first, the types of entities will be identified using the NCBO BioPortal Annotator [11] and S or O will be represented as URI_S and URI_O that are defined by the Annotator API [11]. Second, if entities are no match with existing ontologies, a new concept can be defined for S or O and the conceptual relationship between S or O can be defined by computing the common predecessor of S and O. In this paper, due to the limited space, we will omit the discussion on schema learning.

Property Learning: We are now checking if the objects O in the triplet belong to any of the Classes or Individuals. If they do, the type of the property P is ObjectProperty. If not, it is DataProperty [4].

Triplet Learning: Once the concepts and individuals and properties are identified, the triplets <S,P,O> will be generated with S and O from step i) and P from step ii). These triplets <S,P,O> will be defined as the assertions in both the schema (TBOX) and data (ABOX) [4], as the form of categorized triplets defined from the previous steps.

IV. Implementation

The biomedical application for the proposed framework has been implemented on the Spark parallel engine [16]. For the dataset collection, we used the PubMed API [13]. For the TF-IDF extraction, we implemented using Scala with the Spark MLlib [10]. The NLP, medical term extraction and assertion discovery were implemented using Stanford CoreNLP [8], NCBO BioPortal Annotator [11], and OpenIE [14], respectively. The K-means clustering [3] was implemented using Spark MLlib [10]. The Annotation was implemented using AngularJS with NCBO BioPortal Annotator API [11]. The Ontology construction was implemented on the Spark using OWL API [4] that is a Java API and reference implementation for creating, manipulating, and serializing OWL Ontologies.

V. Results and Evaluation

We randomly selected 30 papers (10 papers for each category as shown in Table I) in three categories such as Diabetes, Obesity, and Heart Disease through the PubMed API [13]. The PubMed API allows us to access publications from more than 27 million citations for biomedical literature from MEDLINE, life science journals, and online books.

From the Assertion Discovery step, we obtained the results as shown in Table I. After the NLP and TF-IDF steps, we have obtained TF-IDF terms (39, 39, and 37 for Diabetes, Obesity, and Heart Disease, respectively) as shown in Table II. Among them, 92%, 76%, 69% are unique to each domain. From the OpenIE step, we obtained 266, 503, and 850 triplets (#Tri) for Diabetes, Obesity, and Heart Disease, respectively. After the mapping the triplets with TF-IDF terms, we found 167, 425, and 703 assertions (#Ass) with a large ratio of unique assertions. The ratios of assertions from the triplets are 66%, 84%, 82% for Diabetes, Obesity, and Heart Disease, respectively. The number of unique subjects (#Sub$_U$), unique predicates (#Pred$_U$), and unique objects (#Obj$_U$) are also reported in Table I. After checking with the NCBO BioPortal Annotator, we found medical terms from subject terms (#Med$_S$) and object terms (#Med$_O$). Some subject or object terms may contain more than one medical term (e.g., 40 medical terms from 21 subjects in the Obesity dataset).

Table III shows the results from the Topic Discovery Component (K-Means). For each domain (as shown in Table III), we found 10 clusters using K-Means Clustering and 10 topics. The sizes of the clusters from K-Means clustering ranged from 0.4% to 45% for Diabetes, 0.5% to 21% for Obesity, and 0.6% to 42% for Heart Disease. Table III shows 6 topics from each domain discovered from K-Means and the number of important assertions.

The medical assertions <S,P,O> for the ontology were dynamically generated from the publications on Diabetes, Obesity, and Heart Disease. From the ontology dynamically generated from 30 medical abstracts, 33 diabetes assertions, 16 obesity assertions, and 21 heart disease assertions have been converted to questions/answers. In our work, we mapped our assertions to the ontological terms in the NCBO BioPortal.

VI. Conclusion

In this paper, we presented a semantic framework that automatically generates an ontology from a large corpus of unstructured text. For this purpose, ontological assertions and topics were discovered from multiple free-text scientific publications on Obesity, Diabetes, and Heart Diseases in PubMed. The pipeline approach in the proposed framework

TABLE I
DATASET OF DIABETES, OBESITY, HEART DISEASE
S: SENTENCE; TRI: TRIPLET; ASS: ASSERTION; $_S$: SUBJECT; $_O$:OBJECT; $_U$: UNIQUE; $_R$: REPRESENTATIVE;

Category	#Abstract	#Word	#Sent	#TF-IDF	#TF-IDF$_U$	#Tri	#Ass	#Ass$_U$	#Sub$_U$	#Pred$_U$	#Obj$_U$	#Med$_S$	#Med$_O$
Diabetes	10	1281	57	39	36	266	169	167	67	83	215	54	191
Obesity	10	638	21	39	30	503	427	425	21	22	99	40	40
Heart Disease	10	1095	44	37	29	850	704	703	49	32	158	45	152

TABLE II
TF-IDF TERMS IN DIABETES, OBESITY, HEART DISEASE

Domain	#TF-IDF	TF-IDF Terms
Diabetes	39	kB, channel, stimulator, cotransporter, Na, Ca2, enhanced, transcription, mutants, dose-dependent, pattern, independent, multiple, urinary, νM, hours, pitfalls, environmental, microscopy, disorders, nickel, detect, racial, plasma, evaluating, concentrations, deposition, cell, minutes, ThT, utilized, impact, proliferator, sampling, FPG, AVS, injured, detected,amyloid
Obesity	39	BMI, males, wage, stimulator, increase, deviation, Na, white, standard, turn, females, two, black, Hispanic, activated, Ca2, cotransporter, channel, venous, stimulates, transcription, enhanced, factor, wages, effect, SGK1, reduces, channels, Vasopressin, further, remodeling, cancer, remains, profile, mechanisms, growth, NFB, restriction, burden
Heart Disease	37	Na, Ca2, ratio, secondary, standard, sources, stimulator, odds, cotransporter, μg, channel, transcription, reduction, enhanced, disease, platelet, SGK1, ensemble, extent, traits, grade, defined, preserved, hyperactivity, phosphatidylinositide-3-kinase, CVD, prediction, failure, acceptor, increases, causes, NFB, 3-phosphoinositide-dependent, expression, disproportionately, hypertension

TABLE III
TOPICS IN DIABETES, OBESITY, HEART DISEASE (AFTER REMOVING SOME IRRELEVANT TOPICS)

Domain	**Topic**	**#Assertions**	**Example of Assertion** $< S, P, O >$
Diabetes	Inhibitory Effect	74	<Inhibitory Effect, WasObservedOn, Mutant HIAPP>
	ThT	195	<ThT, Study, Effect Of Heme On Aggregation>
	Fluid Intake	48	<Fluid Intake, MayRequire, Enhanced Release For Maintenance>
	SGK1	30	<SGK1,IsActivatedBy,Insulin>
	Vein Sampling	11	<Vein Sampling, AreUsedFor, PASubclassification>
	Heart Disease	50	<Heart Disease, Is, One Causes Of Mortality Due To Complications>
Obesity	Childhood Obesity	47	<Childhood Obesity, IsDueTo, Interactions Between Environmental Factors Linked By Epigenetic Mechanisms>
	Cardiovascular Risk Profile	21	<Aim, Evaluate, Risk Profile In Patients With Adrenal Hyperplasia>
	Estimates	6	<I, Estimate, Effect Of Body Mass Index On Wages Across Unconditional Distribution>
	SGK1	13	<SGK1, IsStimulatorOf, Na>
	Studies	9	<Studies,HaveInvestigated,Wall Remodeling>
	FluidIntake	24	<FluidIntake, Require, Enhanced Release Of Antidiuretic Hormone For Maintenance>
Heart Disease	Severe Hypertension	192	Severe Hypertension, IsDiseaseDespite, Optimized Medical Therapy
	SuboptimalFluidIntake	22	<Suboptimal Fluid Intake, Require, Release For Maintenance Of Hydration>
	PlateletHyperactivity	8	<Platelet Hyperactivity, Has, Major Effect On Progression>
	SGK1Expression	5	<SGK1Expression, Favouring, Development Of SGK1 Pathologies>
	Platelets	96	<Platelets, Have, ResponseToProcoagulants>
	Secondary AmmoniumIon	47	<Secondary AmmoniumIon, Responsible, Aerosol>

for sophisticated knowledge discovery and dynamic ontology generation was implemented in Apache Spark. The pipeline includes Stanford CoreNLP for Natural Language Processing, TF-IDF for feature extraction, OpenIE for relation extraction, K-Means clustering for topic discovery and OWL API for ontology generation. The ontology was generated based on the assertions from the pipeline approach and it was also validated by an expert.

REFERENCES

[1] Cimiano, P., and J. Volker. "Text2Onto. Natural language processing and information systems." 10th International Conference on Applications of Natural Language to Information Systems, NLDB. 2005.

[2] Gangemi, Aldo, et al. "Semantic web machine reading with FRED." Semantic Web Preprint (2016): 1-21.

[3] Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." Journal of the Royal Statistical Society. Series C (Applied Statistics) 28.1 (1979): 100-108.

[4] Horridge, Matthew, and Sean Bechhofer. "The owl api: A java api for owl ontologies." Semantic Web 2.1 (2011): 11-21.

[5] Joachims, Thorsten. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. No. CMU-CS-96-118. Carnegie-mellon univ pittsburgh pa dept of computer science, 1996.

[6] Lossio-Ventura, Juan Antonio, et al. "Towards an obesity-cancer knowledge base: Biomedical entity identification and relation detection." Bioinformatics and Biomedicine (BIBM), IEEE Intern. Conference on. 2016.

[7] Ma, Shiwen, et al. "Similarity-based algorithms for disease terminology mapping." Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on. IEEE, 2016.

[8] Manning, Christopher D., et al. "The stanford corenlp natural language processing toolkit." ACL (System Demonstrations). 2014.

[9] Maynard, D., Funk, A., and Peters, W. (2009). SPRAT: a tool for automatic semantic pattern-based ontology population. In Intern. Conference for Digital Libraries and the Semantic Web, Trento, Italy. Citeseer.

[10] Meng, Xiangrui, et al. "Mllib: Machine learning in apache spark." The Journal of Machine Learning Research 17.1 (2016): 1235-1241.

[11] NCBO BioPortal Annotator API https://bioportal.bioontology.org/annotator

[12] Omura, M., N. Sonehara, and T. Okumura. "Practical approach for disease similarity calculation based on disease phenotype, etiology, and locational clues in disease names." Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on. IEEE, 2016.

[13] PubMed Central API https://www.ncbi.nlm.nih.gov/home/develop/api/

[14] Stanovsky, Gabriel, and Ido Dagan Mausam. "Open IE as an Intermediate Structure for Semantic Tasks." (2015).

[15] Velardi, P., Cucchiarelli, A., and Petit, M. (2007). A taxonomy learning method and its application to characterize a scientific web community. Knowledge and Data Engineering, IEEE Transactions on, 19(2):180191.

[16] Zaharia, Matei, et al. "Spark: Cluster computing with working sets." HotCloud 10.10-10 (2010): 95.