A Survey on Bridging Vocabulary Gap in Healthcare with Ontology Support

J.U.Paruvatha Kumari[1] Dr.R.Velumani[2] Dr.S.Nithyakalyani[3]
[1,2,3]Department of Computer Science & Engineering
[1,2,3]K.S.R College of Engineering, Tamilnadu, India

Abstract— Internet is served as a diagnosis tool to facilitate patient-doctor communication. Online health services are deployed to provide remote medical assistance. Community based health services supports automatic disease inference identification for online health seekers. Question and Answer (QA) sessions are suffered with the vocabulary gap and incomplete information. Deep learning scheme is applied to infer the possible diseases using Question and Answer data values. Global learning component is used to mine the discriminant medical signatures from raw features. In local mining raw features and their signatures are updated into the input layer and hidden layer. This paper presents a survey on medical terminology based ontology for improving the relationship in word and detecting diseases.

Key words: Ontology, Global Learning, Local mining, Question and Answer (QA)

I. INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions.

Healthcare industry today generates large amounts of complex data about patients, hospitals resources, disease diagnosis, electronic patient records, medical devices etc. The large amounts of data are a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making. Data mining brings a set of tools and techniques that can be applied to this processed data to discover hidden patterns that provide healthcare professionals an additional source of knowledge for making decisions. The decision rests with health care professionals. Data mining can help healthcare insurers detect fraud and abuse. Healthcare organizations make customer relationship management decisions. Physicians identify effective treatments and best practices, and Patients receive better and more affordable healthcare services.

The clinical and research medical community creates, manages and uses a wide variety of semi-structured and unstructured textual documents. To perform research, to improve standards of care and to evaluate treatment outcomes easily and ideally, in an automated fashion access to the content of these documents is required. The knowledge contained in unstructured textual documents (e.g., pathology reports, clinical notes), is critical to achieving all of these goals. For instance, clinical research usually requires the identification of cohorts that follow precisely defined patient- and disease-related inclusion and exclusion parameters. Biomedical NLP systems extract structured information from textual reports, facilitating searching, comparing and summarization.

II. ONLINE HEALTH RESOURCES

Internet is served as a diagnosis tool to facilitate patient-doctor communication. Online health resources are categorized into two categories. They are the reputable portals and community based health services.

A. Reputable Portals

The reputable portals are run by official sectors, renowned organizations, or other professional health providers. Reputable portals provide up to date health information by releasing the accurate and well-structured health knowledge. WebMD and MedlinePlus are popular reputable portals. WebMD is an American corporation that provides health news, advice, and expertise. It was founded in 1996 by JimClark and Pavan Nigam as Healthscape, later Healtheon, and then it acquired WebMD in 1999 from Robert Draughon to form Healtheon/WebMD. Later, the name was shortened to WebMD. It is primarily known for its public website, which has information regarding health and health care, including a symptom checklist, pharmacy information, drugs information, blogs of physicians with specific topics, and providing a place to store personal medical information. MedlinePlus is an online information service produced by the United States National Library of Medicine. The service provides curated consumer health information in English and Spanish. MedlinePlus provides encyclopedic information on health and drug issues, and provides a directory of medical services. MedlinePlus Connect links patients or providers in electronic health record (EHR) systems to related MedlinePlus information on conditions or medications. A mobile site is also available.

B. Community based Health Services

Community based health services offer interactive platforms to provide Question and Answer (QA) based medical support. HealthTap and HaoDF are the popular community based health services. HealthTap is an Interactive Health company founded by Stanford Graduate School of Business alum Ron Gutman in 2010 to reinvent the way people all over the world take care of their health and well-being. The HealthTap app, both for patients and for doctors, is available across a range of platforms and devices. HaoDF, a Chinese company, owns and operates a healthcare community problem. The company offers medical sites, hospitals, and information to help patients find doctors.

III. LITERATURE SURVEY

Yi Liang Zhao et al., [1] have presented their work on bridging the vocabulary gap between health seekers and providers which delayed the cross-system operability and the interuser reusability. Most of the current health providers organize and code the medical records manually. There is a growing interest to develop automated approaches for medical terminology assignment. To bridge this gap using inventory of medical terminologies local mining and global learning approaches are jointly utilized. Local mining aims to locally code the medical records by extracting the medical concepts from individual record and then mapping them to terminologies based on the external authenticated vocabularies.

They establish a tri-stage framework, which includes noun phrase extraction, medical concept detection and medical concept normalization. Global learning complements the local medical coding in a graph based approach. It collaboratively learns missing key concepts and propagates precise terminologies among underlying connected records over a large collection. The existing techniques can be categorized into two categories: rule-based and machine learning approaches. Rule-based approaches play a principle role in medical terminology assignments. They discover and construct effective rules by making strong uses of the morphological, syntactic, semantic and realistic aspects of natural language. Machine learning approaches build inference models from medical data with known annotations and then apply the trained models to unseen data for terminology prediction. The whole process of the proposed approach is unsupervised and it holds potential to handle large-scale data.

Meng Wang et al., [2] have presented their work using community question answering (cQA) services which gained popularity over the past years. They proposed a novel scheme to answer questions using media data by leveraging textual answers in cQA. It automatically generates a query based on the QA knowledge and then performs multimedia search with the query. Query-adaptive reranking and duplicate removal are performed to obtain a set of images and videos for presentation along with the original textual answer. A graph-based learning process is then formulated based on a regularization framework. Community QA (cQA) has emerged as an extremely popular alternative to acquire information. First, information seekers are able to post their specific questions on any topic and obtain answers provided by other participants. By leveraging area efforts, they are able to get better answers than simply using search engines. Second, in assessment with automated QA systems, cQA usually receives answers with enhanced quality as they are generated based on human intelligence. Third, a tremendous number of QA pairs have been accumulated in their repositories, and it facilitates the conservation and search of answered questions.

M Alfonse et al., [3] proposed Liver cancer ontology in order to give detailed information about liver cancer and also providing semantic information about liver cancer over the web. Ontology was built using the Protégé-owl editing environment. This paper implements web-based Liver Cancer Ontology which consists of five modules: 1) Organizing and scoping. 2) Data collection: Data retrieved from Medicine Net, Cancer.Net, and the National Cancer Institute. 3) Data Analysis: Define classes, properties and create instances. 4) Initial Ontology Development: Ontology is developed using Protégé tool. 5) Ontology Refinement. This ontology can be used by experts or medical researchers.

Liqiang Nie et al., [4] have presented their work, which used content-based approach to automatically predict the search performance. They proposed a query-adaptive graph based learning approach to estimate the relevance probability of each image to a given query. The task is defined as predicting the retrieval effectiveness of a query given a search system and a collection of documents. The probabilities should be close to the ranking-based relevance probabilities. A graph is constructed based on the search results of a query, where vertices are the images and edge weights indicate the pair-wise similarities

Ibrahim F. Moawad et al., [5] proposed ontology for viral hepatitis in order to help physicians by using developed ontology in knowledge sharing, reasoning and exploiting. Methodology consists of three phases: VH Ontology Extraction, VH Ontology Validation and VH Ontology Representing in OWL. In VH Ontology Extraction: Related knowledge is extracted from domains. In order to build ontology bottom up approach is being followed .Bottom up approach means all the information is extracted by from the lowest level of granularity. In VH Ontology Validation: Domain experts validate the review of results. In VH Ontology Representing in OWL: By using protégé editor tool this ontology is being constructed.

Pakhomov et al., [6] attempted to increase the coding performance by combing the advantages of rule-based and machine learning approaches. It represents Autocoder, an automatic encoding system implemented at Mayo clinic. Autocoder integrates example-based rules and a machine learning module using Nave Bayes. However, this combination is loosely coupled and the learning model cannot integrate heterogeneous suggestion, which is not a valuable choice for the community based health services. Beyond medical domain, several prior efforts of corpus alignment and gap linking have been dedicated to other verticals, and return a precise text that provides the answers.

Maja Hadzic et al., [7] proposed Ontology support for Human Disease Study. The Proposed solution incorporates a model for Generic Human Disease Ontology. It contains information about human diseases. Generic Human Disease Ontology consists of four branches: types, phenotype, causes responsible for that disorder and treatments. Ontology helps to provide Specific Human Disease Ontologies. Types Branch: Describes different types of disorders. Phenotype Branch: Describes about symptoms of a disease. Treatment: An overview of all the treatments for that particular disease. Generic Human Disease Ontology is constructed using DOGMA Modeler tool. Advantages are computer based ontology supports the scientist in collecting information about disorders, it improves shared knowledge efficiency and effectiveness, ontologies allows distributed resources which are autonomous and heterogeneous to function in a worldwide environment. This help to split a big task between different teams effectively.

IV. METHODS

To prevail over these limitations, we propose a novel scheme that is able to code the QA pairs with corpus-aware terminologies. As illustrated in Figure 1, the proposed method consists of two jointly augmented components that is, local mining and global learning.

A. *Local Mining Approach*

Medical concepts are referred to medical domain specific noun phrases and Medical terminologies are allude to as authenticated phrases by well-known organizations that are used to accurately describe the human body and associated components, circumstances and processes in a science-based method. It establishes a tri-stage framework. First to extract noun phrase then identify medical concept and finally it normalize the medical concepts to terminologies.

1) *Noun Phrase Extraction*

To extract entire the noun phrases, we originally assign part-of-speech tags to each word in the given medical record by Stanford POS tagger. And then pull out sequences that match a fixed pattern as noun phrases.

The regular expression can be possibly interpreted as follows. The noun phrases should consist of zero or more adjectives or nouns, followed by an optional group of a noun and a preposition, followed again by zero or more adjectives or nouns, followed by a single noun. A sequence of tags matching, this pattern ensures that the corresponding texts make up a noun phrase. For e.g., the following complicated sequence can be extracted as a noun phrase: “ineffective treatment of terminal lung cancer”. In addition to simply pulling out the phrases, and we also do some simple post processing to link the variants together, such as singularizing plural variants.

2) *Medical Concept Detection*

In this stage aims to differentiate the medical concepts from other general noun phrases. Inspired by the efforts, it consider that concepts are relevant to medical domain occur frequently in medical domain and seldom in non-medical ones. The concept entropy impurity (CEI) is a similarity measure for domain relevance of a concept. In that impurity used c as concept, D1 and D2 represents our medical corpus and a general-domain corpus. P(Di j c) denotes the probability that a concept is related to a specified domain Di.

3) *Medical Concept Normalization*

Although medical concepts are explain as medical domain-specific noun phrases, we cannot provide that they are standardized terminologies. For e.g., “birth control”, that is recognized as a medical concept by local mining approach, but it is not an authenticated terminology. Rather than, it maps into “contraception”. Hence, it is necessary to normalize the detected medical concepts according to the external appropriate standardized dictionary and this normalization is the key to linking the vocabulary gap.

Right now, there exist several authenticated vocabularies, including ICD7, UMLS, WordNet and SNOMED CT. These medical and clinical terminologies were invented in different times by different associations for multiple purposes. For e.g., ICD, in general it is used for external reporting requirements. In this work, we use WordNet because it provides the general terminologies for the electronic health record and formal logic-based hierarchical structure.

B. *Global Learning Approach*

Global learning is an important method, including local approach, and attempted to map the QA pairs directly to the entries in external dictionaries without any pruning. This method generally presents problems since the external dictionaries naturally cover relatively comprehensive terminologies and are far beyond the vocabulary scale of the given corpus. It may result in the deterioration in coding performance in conditions of efficiency and effectiveness. The problem is caused by the over-turned scope of vocabularies, which may take in unpredictable noises and make the precise terminology selection challenging. As a result, a corpus aware vocabulary terminology is naturally constructed by local mining approach, which can be used as terminology gap for further learning.

Let Q = {q1; q2;…..qn} and T = {t1; t2; ….tm} respectively represent a repository of QA pairs and their connected respectively denotes a repository of medical records and their connected locally mined terminologies. The target of global learning is to learn appropriate terminologies from the global vocabulary space T to interpret each medical record q in Q. Along with existing machine learning methods; graph-based learning achieves promising performance. In that paper, we also explore the graph-based learning model to accomplish terminology selection task, and expect this model is able to simultaneously consider various heterogeneous cues, including the medical record content analysis, terminology-sharing networks, and the inter-expert as well as inter-terminology relationships. We will first propose relationship identification and then we explain in detail, how to use our proposed model to associate the underlying connected medical records. Then, we present the optimal solution for learning model followed by the label bias estimation.

1) *Relationship Identification*

The inter-terminology and inter-expert relationships are not intuitively implied from medical records, so we call them as implicit relationships. This subsection purpose is to introduce how to discover these kinds of relationships.

a) Inter-Terminology Relationship

The medical terminologies in WordNet are organized into acyclic taxonomic (is-a) hierarchies. For e.g., “viral pneumonia” is “infectious pneumonia” is-a “pneumonia” is-a “lung disease”. Terminologies may have multiple parents. For e.g. “infectious pneumonia” is a child of “infectious disease”. The inter-terminology hierarchical relationships are semantically capture well-defined ontology.

b) Inter-Expert Relationship

The inter-expert relationships will be viewed stronger if the experts are professionals in the same or related specific medical domain. This is returned by their historical data, i.e., the number of questions they have co-answered.

2) *Probabilistic Hypergraph Construction*

The graph based learning models can be broadly categorized into simple graph-based and hypergraph based approaches, they are built on a graph where vertices are samples. The simple graph conveys the pair-wise relationship of vertices and overlooks the relations in higher orders, which are sensitive to the radius parameter used in similarity calculation. As compared to simple graph, hypergraph contains the summarized local grouping information by allowing each hyperedge to connect more than two vertices simultaneously. Meanwhile hyperedge types and weights can be empirically set according to certain rules, and they can be heterogeneous to fuse comprehensive and diversified sources. Taken mutually, hypergraph-based learning partially fits task of terminology selection via integrating multi-faceted information cues, except considering the inter-terminology hierarchical relationship.

3) *Global Learning Optimization*

Global learning optimization has been defined the hypergraph based framework for the global terminology learning that contains three objectives, and we aim to formulate each objective in details and derive a solution to this optimization problem. The ideas to formulate these three objectives are as follows. The first objective should guarantee that the relevance probability function is continuous and smooth in semantic gap. It means that the relevance probabilities of semantically similar medical records should be close to each other. The second objective is ensured by the empirical loss function, which forces the relevance probabilities to approach the initial roughly appropriate relevance scores. These two implicit constraints are widely adopted in reranking oriented methods. The third objective encourages the standards of medical records, which are connected by hierarchical structured terminologies, should be similar to each other.

4) *Pseudo Label Estimation*

The idea of empirical loss term is to ensure the learnt relevance probabilities between terminologies and medical records are not far away from the initial roughly estimated relevance scores. Another method to decrease the size of the hypergraph is by pre-clustering the medical records during the data collection stage into several subgroups, and the hypergraph-based learning is conducted within each cluster. Each cluster contains the semantically close medical records, which are inter connected and most probably share the same vocabulary.

C. *Ontology*

The use of ontologies in medicine is mainly focussed on the representation and (re-)organization of medical terminologies. Physicians developed their own specialized languages and lexicons to help them store and communicate general medical knowledge and patient-related information efficiently. Such terminologies, optimized for human processing, are characterized by a significant amount of implicit knowledge. Medical information systems, on the other hand, need to be able to communicate complex and detailed medical concepts (possibly expressed in different languages) unambiguously. This is obviously a difficult task and requires a profound analysis of the structure and the concepts of medical terminologies. But it can be achieved by constructing medical domain ontologies for representing medical terminology systems.

Ontology-based applications have also been built in the field of Medical Natural Language Processing.

1) *Benefits of ontology*

– Ontologies can help build more powerful and more interoperable information systems in healthcare.
– Ontologies can support the need of the healthcare process to transmit, re-use and share patient data.
– Ontologies can also provide semantic-based criteria to support different statistical aggregations for different purposes.
– Possibly the most significant benefit that ontologies may bring to healthcare systems is their ability to support the indispensible integration of knowledge and data.

V. Results

A. *Document Retrieval*

The document retrieval process can retrieve locally stored documents; its remote facility retrieves the relevant documents from medical websites using the Google search service. These medical websites are sorted from the previously defined medical website classification. This medical website classification is performed before the real-time execution of the google search engine and consists of defining the different medical website classes where the system can retrieve the medical documents. Document retrieval engine can start retrieving those relevant documents from medical websites whether there exists or not the association between the searched generic question and the medical websites. When the treated generic question has been related to at least one medical websites class then the Google search engine retrieves the relevant documents according to the question keywords in these medical websites.

B. *Relevant Passage Selection*

Relevant Passage Selection process consists of extracting the sentences from these medical documents that could answer questions of the user easily. These sentences are extracted by applying a technique based on comparing the question keywords in the documents and, those sentences that at least contain a question keyword are extracted from the document and are evaluated by the next Answer Extraction module that decides if the sentence correctly answers the user question.

C. *Answer Extraction*

Answer to any question asked by the user is extracted with the help of answer extraction process which extracts the answer by analyzing the sentences extracted by the previous relevant passage selection module. This process is performed by applying the following steps: the first one consists of inferring the logic form of the sentence and identifying the main verb in this logic form; the following step is to verify if this main verb belongs to the set of verbs that can answer the generic question; the third step is the acknowledgment of the medical entities in the logic form; the next step is of comparing if the medical entities searched as the answer is found in the logic form; and finally, the last step is the analysis of the predicates that relate the answer of the candidate, the main verb and the rest of the medical entities in the answer form. This module produces Ranking of Answers. The verb can distinctively relate two medical entities considering this feature as a direct link.

VI. Conclusion

A review of the research work for vocabulary gap in Health care is discussed in this paper. The Local-Global learning approaches are efficient approaches for reducing the vocabulary gap in health care domain. For more efficient to integrate the constructed ontology to identify the concept relationship based discriminant features for each specific diseases. And we will also flexibly organize the unstructured medical content into user needs-aware ontology by leveraging the recommended medical terminologies.

REFERENCES

[1] Liqiang Nie, Yiliang Zhao, Mohammad Akbari, Jialie Shen, and T Chua. “Bridging the vocabulary gap between health seekers and healthcare knowledge”, In IEEE Transactions, 2014.

[2] Beyond text qa: Multimedia answer generation by harvesting web information. L. Nie, M. Wang, Y. Gao, Z.-J. Zha, and T.-S. Chua. IEEE Transactions on Multimedia, 36(4):8509-8522, 2013.

[3] Marco Alfonse, Mostafa M. Aref, Abdel-Badeeh M. Salem, ―Ontology- Based Knowledge Representation for Liver Cancer‖, Electronics Proceedings of the Internationall eHealth Telemedicine and Health ICT Forum for Educational, Networking and Business, Pages 840-844,2012.

[4] Oracle in image search: A content-based approach to performance prediction. L. Nie, M. Wang, Z.-J. Zha and T.-S. Chua. ACM Transactions 2012.

[5] Ibrahim F. Moawad, Galal AL Marzoqi, Abdel-Badeeh M. Salem, ―Building OBR-based OWL Ontology for Viral Hepatitis‖, Electronics Proceedings of the Internationall eHealth elemedicine and Health ICT Forum for Educational, Networking and Business, pp.821-825, 2012.

[6] Pakhomov, Serguei VS, James D. Buntrock, and Christopher G. Chute. “Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques”, Journal of the American Medical Informatics Association 13.5 (2006): 516-525.

[7] Maja Hadzic, Elizabeth Chang, ― Ontology-based support for Human Disease Study‖, International Conference on System Sciences-2005.