Ontological Approach for Knowledge Extraction from Clinical Documents

1st Raxit Goswami
Research Lab
ezDI Inc.
Louisville, USA
raxit.g@ezdi.com

2nd Vatsal Shah
Research Lab
ezDI Inc.
Louisville, USA
vatsal.s@ezdi.com

3rd Nehal Shah
Research Lab
ezDI Inc.
Louisville, USA
neil@ezdi.com

4th Chetan Moradiya
Research Lab
ezDI Inc.
Louisville, USA
chetan.m@ezdi.com

***Abstract*—In clinical NLP (Natural Language Processing), Knowledge extraction is a very important task to develop a highly accurate information retrieval system. The various approaches used to develop such systems include rule-based approach, statistical approach, shortest path algorithm or hybrid of these approaches. Accuracy and coverage are the most important parameters while comparing different approaches. Some methodologies have good accuracy but low coverage and vice-versa. In this paper, our focus is to extract domain relationships, for example to extract the relationship between 'Disease' and 'Procedure' or 'Symptom' and 'Disease' etc. from the clinical documents using three different approaches. These three approaches are i) Statistical ii) Shortest Path iii) Shortest Path Using Body System. All three approaches use our in-house existing NLP system to extract entities from the un-structured documents. The Statistical approach applies a probabilistic algorithm on clinical documents, whereas the Shortest Path algorithm uses the Ontological knowledge base for the hierarchical relationship between entities. This Ontological knowledge base is built upon the curated Unified Medical Language System (UMLS). For the Shortest Path Using Body System approach, we have used the domain relationship as well as hierarchical relationship. The output of these approaches is further validated by a domain expert and this validated relationship is used to enrich our ontological knowledge base. We have presented the details of these approaches one-by-one along with the comparative results of these approaches. We finally go through the analysis of the result and conclude on further work.**

***Index Terms*—Knowledge Extraction, Clinical information retrieval, Relationship Extraction, Clinical Document, Medical knowledge base, Ontology, Clinical NLP (Natural Language Processing)**

I. Introduction

Nowadays, the clinical knowledge plays an important role in the medical field. Clinical knowledge is growing quickly. This large scale clinical data can also be useful in many clinical tasks like improving access to care, patients predictions, search engines such as PubMed [1] and CISMeF [2], predictive analytics, and clinical document improvement [3], etc. Clinical knowledge is in the form of concepts like symptoms, diseases, procedures, anatomical structure, medicines, findings, medical devices, and body measurements. These concepts are also used to identify relationships in medical documents. So, to extract medical entities from unstructured data and the right relationship between these entities is a very challenging task.

Some research work already has been done in this field. MeTAE [4] is a rule-based system which extracts annotated medical entities and relationships from medical documents. This approach depends on linguistic patterns and domain experts. The limitation in this system is that it requires to have a specific qualifier for the target relation to obtaining more focused pattern construction in the corpus. MetaMap [5] is another tool which maps medical text to UMLS concepts. MetaMap is developed to identify clinical findings, molecular binding, drugs, genes and relationships between them from MEDLINE [6] citations or clinical reports. MetaMap has some limitations as it detects wrong medical entities for chemical names, abbreviations, and acronyms. Another limitation is that this tool detects some general words as medical entities and also categorizes the same words into two different concept types. A classic disadvantage of pattern-based or rule-based methods is the expensive cost needed to obtain a good recall.

There are certain systems which have also used a statistical approach to extract relationships between entities. The one such system [7], uses a statistical approach in which the author first performed vision-based web entity extraction and then used a statistical algorithm to find the relevant relationships between extracted web entities. As they are using a web database for this task the accuracy of entity extraction is quite good but, results of relationship extraction are not up to the mark. DeepDive [8] is a system which performs knowledge-base construction (KBC) from hundreds of millions of web pages that uses statistical learning and inference. The inference model works well with the only large amount of data. There has been good research work on relationship extraction using semantic-web technology [9]–[11]. They are proposing relationship extraction from EHR (Electronic Health Record) where they have used knowledge graph to traverse and infer to extract meaningful insights. Another system [12] has described the ontology-driven approach for knowledge extraction where they are limited to small ontology and have less f-measure.

In this paper, we have explained three different approaches to extract medical concepts and assign the relation between these concepts. To extract the right medical entities from unstructured documents, we have used our in-house existing NLP [13] system. NLP converts unstructured medical documents into structured information and maps the clinical concept to

a unique identifier and its type using curated UMLS [14]. The basis of curated UMLS is also UMLS [15], [16] but we have done preprocessing to remove ambiguity and redundant data. The output of our NLP system and ontology is used for relationship extraction.

II. Overview of Ontology

Ontology is a set of concepts and categories in a subject area or domain that shows their properties and the relations between them. An ontology consists of concepts and relationships between these concepts. The concepts are real-world entities that are mapped to appropriate classes. Our ontology is built using many sources like curated UMLS, verified medical articles, and books. It is made up of 1.7 million concepts which are mapped to the major 7 class. The classes that are present in the ontology are mentioned below:

1) Medications
2) Procedure
3) Anatomical Structure
4) Symptoms
5) Disorder
6) Body Measurement
7) Findings

There are two types of relationships in an ontology (i) Hierarchical Relationship (ii) Domain Relationship. The total number of relationships between concepts are 2.2 million. Our Ontology is very rich in hierarchical relationships which describe parent/child of any concept and it is defined as "is_subclass_of" in our ontology. The "Domain Relationships" are the key components of an ontology. However, it significantly lacked the domain relationships between these concepts. Out of 2.2 million relationships, ontology had only 0.16 million domain relationships which means that there is a significant knowledge gap in the ontology. The task was to come up with an efficient algorithm to find these missing domain relationships, with minimal human effort. The most important domain relationships in our ontology are shown in Table I.

TABLE I
Domain Relationships with Description

Domain Relationship	Description
is_Symptom_Of	Relationship between diagnosis and symptoms
is_Procedure_Of	Relationship between diagnosis and procedures
is_Medication_Of	Relationship between diagnosis and medications
has_Location	Relationship between diagnosis, procedures, symptoms to an anatomical structure
has_Finding	Relationship between diagnosis and finding

For this, The clinical document is first parsed using our NLP system and an XML document is generated. The concepts are extracted from the XML document using the relevant concept type. The focus should be on extracting only the relevant concepts from the document. If we are interested in finding the "is_Symptom_Of" relation then we need to extract the symptoms and disorders that are mentioned in the document. Other details like the medications that the patient is taking and the medical procedures that the patient has undergone are irrelevant in this case. The focus should be on the kind of relationship that needs to be captured in the ontology. Once that is done then the relevant concepts can be extracted from the document. Once the relevant concepts are extracted, then the relationships can be extracted, and these relationships need to be validated by the domain experts before they are pushed into the ontology. For example, in the "is_Symptom_Of" relationship, the concepts that would be relevant are the symptoms and disorders. Suppose there are 50 symptoms and 100 disorders in the document, then there are 5000 potential relationships, the majority of which will not be correct. Hence if take this approach, we will be wasting a major portion of the domain expert's time who will be validating these relationships. The task was to develop an algorithm, which will suggest plausible relationships that will decrease the burden of the domain experts.

III. Approach

The solution to the problem mentioned above lies in the three different approaches that will be discussed in detail in this report. Each approach serves a certain purpose and has varying performance measures. The task of the approach is to identify related entity pairs and identify the keywords that indicate the relationships. The three approaches are as follows:

1) Statistical
2) Shortest Path
3) Shortest Path Using Body System

A. Statistical Approach

We define a probabilistic model which is applied to clinical documents to identify related entity pairs. To build a probabilistic model we have used two coefficients (i) Jaccard Coefficient and (ii) Percentage. The count of the first Entity is defined as X and the second entity is defined as Y. The Co-occurrence of both entities is defined as CC. The equation of the Jaccard Coefficient is defined as below:

$$\frac{\text{Co-occurrence Count(CC)}}{\text{(Sum of Individual Entity Count(X+Y)) - Co-occurrence count(CC))}} \times 100 \quad (1)$$

Similarly, The equation of Percentage is defined as below:

$$\frac{\text{Co-occurrence Count(CC)}}{\text{Min(First Entity Count(X),Second Entity Count(Y))}} \times 100 \quad (2)$$

To test our approach we have used 100 clinical documents from our corpus, consist of different varieties of documents like radiology reports, cardiology reports, etc. (i.e based on service line). These documents are physicians notes from

hospitals and clinics of the US. The documents have a large part as unstructured text and also some tables and template sentences.

The steps involved in this approach are as follows:

1) Categorized the clinical documents based on category (eg: cardiology, urology, etc.)
2) Extract the concepts from relevant sections(eg: Impression and Plan, Diagnosis, etc.)
3) Count the individual occurrence of each concept and the co-occurrence count of each pair of concepts.
4) Use the JACCARD coefficient and percentage individually and together to find the probability of two concepts to be related to each other.

The results of this approach are shown in Table II and Table III. It Indicates that precision is 64.44% for Jaccard Coefficient=1 which is good for baseline and there is scope for improvement. This approach is good if we are building a knowledge base from scratch. It does not leverage the relationships that already exist in the ontology. However, It requires a very large corpus and corpus has to be categorized based on the category.

TABLE II
OVERALL PRECISION USING THE JACCARD COEFFICIENT

Jaccard Coefficient	Total Suggestions	Correct Suggestions	Wrong Suggestions	Precision (%)
1	45	29	16	64.44
0.5 to 1	78	34	44	43.58

TABLE III
OVERALL PRECISION USING THE PERCENTAGE AND JACCARD COEFFICIENT

Jaccard Coefficient and Percentage	Total Suggestions	Correct Suggestions	Wrong Suggestions	Precision (%)
>=0.25 and 100	74	42	32	56.75
0.2 to 0.25 and 100	31	14	17	45.16

B. Shortest Path Approach

This approach leverages the relationships that already exist in the ontology. To extract relationships between two entities we need to traverse from one entity to another entity using the shortest path in the ontology. Once we reached to destination node by the shortest path we can infer that these two entries are related to each other.

The steps involved in this approach are as follows:

1) Parse the clinical document to generate the XML document.
   a) Filter the XML document based on sections.
   b) Suggest relationships between concepts that are in the same sentence.
2) Compute the shortest path between the concepts that are extracted from the document. We have tried using multiple values of the shortest path and generate the output.
3) Validate the suggested relationships by the domain experts.

Similarly, to test this approach we have used 100 clinical documents from our corpus and the results are shown in Table IV. The precision is very high for the lower shortest path and it keeps decreasing respect to shortest path length. The advantages of this approach over the previous are that it does not require a large corpus to extract relationships and corpus need not be categorized based on the category.

TABLE IV
OVERALL PRECISION USING THE SHORTEST PATH APPROACH

Shortest Path	Total Suggestions	Correct Suggestions	Wrong Suggestions	Precision (%)
1	1	1	0	100
2	14	10	4	71.42
3	16	10	6	62.5
4	50	27	23	54
5	69	37	32	53.62
6	33	12	21	36.36

The limitation with this approach is that here we need to define the nodes (i.e. shortest path length) which we need to traverse to check the connection between two entities in the ontology. However, as it is evident from the result that as the shortest path length is increased, the accuracy decreases. Conversely, if the length is decreased, the coverage is very low as the algorithm misses some of the potential relationships. Also, sometimes the high threshold of the shortest path gives a relationship without context, which may require more human efforts for validation.

C. Shortest Path Using Body System

To overcome the limitation of the above approach, we have used 'Domain Relationship' along with 'hierarchical relationship'. The intuition behind this is that two related concepts in the medical domain always relate to the same part of the body. For example, a disorder and a symptom affect the same part of the body, or a medical procedure is done on the same body organ which is affected by a particular disorder. Our Ontology provides information about the 'superclass' and/or 'subclass' (Hierarchy Relationship) of the entity, and 'has_Location' (Domain Relationship) to body system which directly and indirectly helps to identify the relationship amongst entities.

To identify the probable relationship between two entities, the hierarchy of anatomical structure is validated manually by a domain expert. Its validated in a way where the anatomical structure of two different entities who share a similar superclass till the common body region/system/organ site and hence they are found to have a probable relationship among each other.

To illustrate it further, we take a sample of the hierarchy of anatomical structure where validated classes are highlighted as

shown in “Figure 1”. It is important to decide which superclass of the entity should be considered for the relationship. For example, In the case of “Brain Structure”, it has a superclass which combines the brain and spinal cord region which may not necessarily suggest a relationship amongst each other and it is not ideal to go up to that level. Suppose, if “Brain and Spinal cord Structure” superclass is considered, then two entities like “spinal fracture” and “brain tumor” both will be predicted as related entities as they share the same superclass which is a false-positive result.

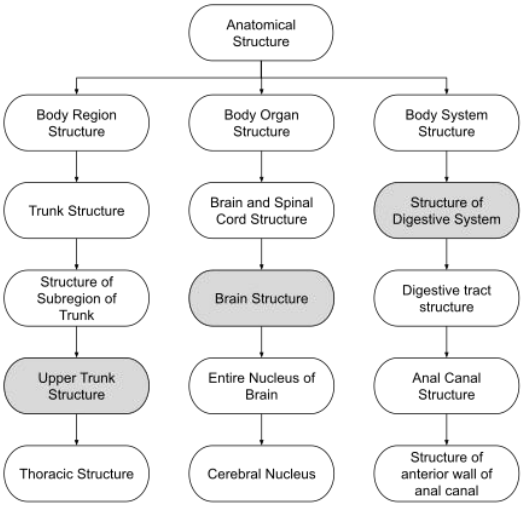

Fig. 1. Validation in hierarchy of Anatomical Structure.

The steps involved in this approach are as follows:

1) Parse the clinical document to generate the XML document.
2) Extract all the relevant concepts from the XML document.
3) Find the body system which is being affected by these concepts. This is done by leveraging the has_Location relationship that is present in the ontology.
4) Find the least common parent of the body systems that are being affected by the two concepts.
5) Check if the relationship is plausible, if yes forward the relationship to the domain expert for validation.

To explain how the system identifies the relationships from the clinical document, we take an example of **Chest pain** and **Bacterial pneumonia** where both entities are present in the same clinical document. As we can see in “Figure 2” chest pain occurs in the “chest” region and the chest has a location relationship with “Thoracic structure.” In the same way, Bacterial pneumonia has a location of “Lung structure” which is a subclass of “Thoracic viscus” which has a superclass “Structure of compartment of thorax”. This entity has a superclass “Thoracic structure”. So, by using the reference of common anatomical structure (i.e. “Upper Trunk Structure”), we can predict that two entities are related to each other as they share a common anatomical structure. We analyzed certain examples and with the help of domain experts, we decided to relate the entities upto the level of the similar anatomical structure. The advantage here is that we dont require to specify the length of the path to get the desired result.

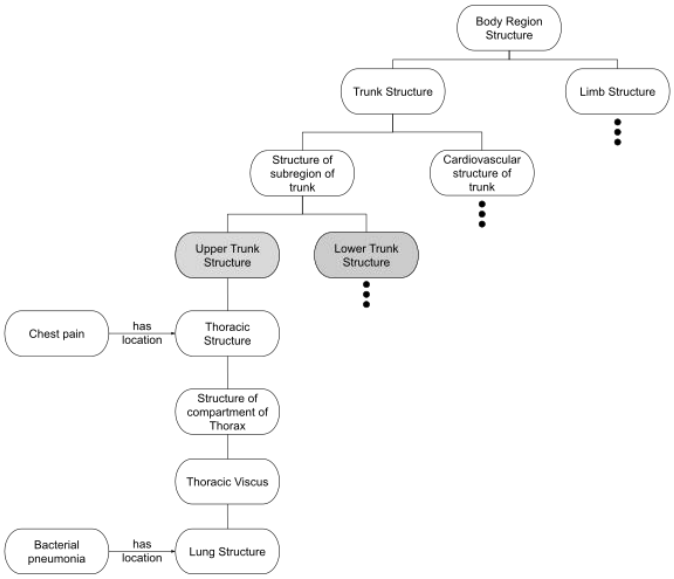

Fig. 2. An example of Relationship between ‘Chest pain’ and ‘Bacterial pneumonia’ in the ontology.

Unlike previous approaches, to test this approach we have used 1,000 clinical documents from our corpus, consist of different varieties of documents (i.e based on service line). After applying our approach, we have found a total of 4,346 unique relationships and randomly 1,000 relationships are given to domain experts to validate. The results of this approach are shown in Table V. The precision is 89.3% which is very high compared to all previous approaches.

TABLE V
OVERALL PRECISION USING THE SHORTEST PATH USING BODY SYSTEM

Total Documents	1,000
Total Unique Suggestions	4,346
Validated Suggestions	1,000
Correct Suggestions	893
Wrong Suggestions	107
Precision(%)	89.3

From the results, it is evident that this approach gives the best output among the above two approaches and also overcome its limitations. Using this approach we do not need to categorize the corpus based on category. The parsed documents need not be filtered based on sections. Moreover, we can process one file at a time instead of having a huge corpus while extracting relationships. After analyzing the false negatives result, we found that some entity has wrongly mapped in the ontology and this can be corrected with the help of domain experts. This approach misses the possible

relationship between two entities when there is no relation to body structure in the ontology. So, this can be the only limitation of this approach where we need to have accurate domain relationships in the ontology and this can be easily achieved with the help of domain experts.

IV. Conclusion

This paper presented an ontological approach for knowledge extraction from clinical documents. It is evident that the last approach is better in terms of accuracy and coverage. It overcomes all the limitations which our previous approaches have. This meets our initial objective, which is to have high precision and high coverage in relation extraction. Comparing to another existing system, our approach establishes good results in precision. The results collected on a real test corpus represent the effectiveness of our approach and its advantages. In transient viewpoint, our aim is to contemplate the false negatives to improve our system and ontology. With the help of domain experts, we can get insights about how we can fine-tune this system. We also intend to try with different open-source clinical corpse to validate our system.

References

[1] PubMed, "Pubmed search engine," 1996, http://www.pubmed.com, Last accessed on 2019-02-08.

[2] CISMeF, "Cismef search engine," 1998, http://www.chu-rouen.fr/cismef, Last accessed on 2019-02-10.

[3] V. Shah, R. Goswami, V. Kumar, B. Shah, and H. Shah, "Automated clinical documentation improvement," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2018, pp. 1544–1547.

[4] A. B. Abacha and P. Zweigenbaum, "Automatic extraction of semantic relations between medical entities: a rule based approach," *Journal of biomedical semantics*, vol. 2, no. 5, p. S4, 2011.

[5] A. R. Aronson, "Effective mapping of biomedical text to the umls metathesaurus: the metamap program." in *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2001, p. 17.

[6] W. Pratt and M. Yetisgen-Yildiz, "A study of biomedical concept identification: Metamap vs. people," in *AMIA annual symposium proceedings*, vol. 2003. American Medical Informatics Association, 2003, p. 529.

[7] Z. Nie, J.-R. Wen, and W.-Y. Ma, "Statistical entity extraction from the web," *Proceedings of the IEEE*, vol. 100, no. 9, pp. 2675–2687, 2012.

[8] F. Niu, C. Zhang, C. Ré, and J. W. Shavlik, "Deepdive: Web-scale knowledge-base construction using statistical learning and inference." *VLDS*, vol. 12, pp. 25–28, 2012.

[9] S. Perera, C. Henson, K. Thirunarayan, A. Sheth, and S. Nair, "Semantics driven approach for knowledge acquisition from emrs," *IEEE journal of biomedical and health informatics*, vol. 18, no. 2, pp. 515–524, 2013.

[10] A. Sheth, I. B. Arpinar, and V. Kashyap, "Relationships at the heart of semantic web: Modeling, discovering, and exploiting complex semantic relationships," in *Enhancing the Power of the Internet*. Springer, 2004, pp. 63–94.

[11] S. Perera, C. Henson, K. Thirunarayan, A. Sheth, and S. Nair, "Data driven knowledge acquisition method for domain knowledge enrichment in the healthcare," in *2012 IEEE International Conference on Bioinformatics and Biomedicine*. IEEE, 2012, pp. 1–8.

[12] N. M. Matasyoh, G. Okeyo, and W. Cheruiyot, "Ontology-driven approach for knowledge sharing and retrieval," *International Journal of Computer Science Issues (IJCSI)*, vol. 13, no. 4, p. 59, 2016.

[13] P. Pathak, R. Goswami, G. Joshi, P. Patel, and A. Patel, "Crf-based clinical named entity recognition using clinical nlp," in *Proceedings of International Conference on Natural Language Processing*, 2013.

[14] V. Shah, B. Shah, R. Goswami, S. Kumar, and C. Moradiya, "Creation of unambiguous centralized knowledge base from umls metathesaurus," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2018, pp. 1270–1276.

[15] National Library of Medicine (US), "Umls 2006aa documentation," 2006, https://www.nlm.nih.gov/research/umls/2006AA_umls_documentation.pdf, Last accessed on 2019-01-10.

[16] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.