Proceedings of the 5th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'19)
Lisbon, Portugal – August, 2019
Paper No. MVML XXX (The number assigned by the OpenConf System)
DOI: TBA

Mining Symptom and Disease Web data with NLP and Open linked Data

Hong Qing Yu
School of Computer Science and Technology
University of Bedfordshire
Luton, United Kingdom
Hongqing.yu@beds.ac.uk

Abstract - Machine Learning (ML) technologies in recent years are widely applied in various areas to assist knowledge gaining and decision-making on healthcare. However, there is no reliable dataset that contains semantic structured knowledge on symptom and disease enable to apply advanced machine learning algorithms such clustering or prediction. In this paper, we propose a framework that can extract data from web with apply Natural Language Processing (NLP) process and semantic annotation to create Open Linked Data (OLD) based knowledge graph. At the end, the knowledge graph can be used for ML algorithms and graph oriented Deep Learning techniques.

Keywords**:** Machine Learning, Open Linked Data, Natural Language Processing, Knowledge Graph

1. Introduction

Medical or healthcare related information on the Internet has grown enormously in recent time. On the one hand, the number of internet users turning to search health-related information online continues to increase according to recent study from the Pew Internet & American Life Project in July 2018. On the other hand, various machine learning research works from the past tend to use the information they captured from online resources such as social media, communication forum and many other resources to create AI supported healthcare recommendation applications. The research results are very encouraged that these AI applications can provide helpful tips or even pre-diagnostic advices based on very simple datasets e.g. condition and symptom relational datasets [1]. However, they are majorly based on one specific condition information e.g. heart disease condition, which request to have strong pre-prediction to a certain disease firstly enable getting AI confirmation. To fix this issue recent researchers suggested to use full common disease-symptom dataset to predict or classify disease conditions as a whole rather than one single disease. There are many different features between single condition prediction and multiple conditions prediction by ML. The difference is that the common condition dataset is more about general symptoms rather than specific and detailed health markers. The other important difference is that the data for common conditions are only based on the relation between symptoms and conditions with rare other information. Thus, there is no information about individual case that contains profile information such as age and sex. In addition, there is no final condition concluded with target indicators for ML, e.g. yes (1) or no (0). Therefore, the common condition ML is more difficult and complex comparing to single condition ML. These differences indicate a challenge of create a disease-symptom knowledge graph with machine understandable relations.

Therefore, we propose a semantic feature enrichment process to add DBpedia terms (URI links) together with natural language processing tokens to analyse symptom descriptions for 215 conditions. The data is automatically crawled from UK NHS website. With enriched symptom tokens, the unsupervised clustering algorithm can be applied to provide recommendation rankings of conditions to symptom inputs. The rest of the paper is organised as:

Related work about current symptom based condition prediction research and disease knowledge representation will be discussed in Section 2. The proposed semantic enrichment process will be illustrated in Section 3. The experimental results on exploring the dataset in Section 4. Finally the conclusion will be in Section 5.

2. Related Work

There are many machine learning algorithms that have already been applied to use datasets available online (Kaggle platform[i]) or their own research datasets.

The work [2] presents an analysis framework which compares different types of unsupervised clustering ML algorithms with latent class. The algorithms include k-means, Birch, Spectral-Clustering, Hierarchical Agglomerative Clustering (HAC) and k-Modes. The conclusion is that k-Modes has the best performance to identify the most separable clusters.

The paper [3] suggests a symptom based disease prediction system architecture by combining four-type of ML algorithms together sequentially to produce highly accurate result. However, the work does not analyse any possible use case or dataset. The major problem is that the accuracy and efficiency are badly balanced. Meanwhile the suggestion is not aware that different ML algorithm requires different data feature engineering.

A NLP work [4] develops an expertise system to help doctor for diagnoses. The system contains a NLP process try to analyse free text information from user and mapping the information to a fix set of symptoms and feature definitions (see Table 1). Finally, the Naïve Bayes algorithm is applied to classify the symptoms to the conditions. The work claimed that the accuracy of classification rate can reach 92.2819%. However, only 4 conditions are tested in the system and 300 records are used. In addition, the dataset is not available to be examined by other researchers to understand the usage of time related features e.g. cold < 7 days but no similar information about cough or other symptoms.

The other related research work [5] also proposes a decision tree based general disease prediction system. However, this process does not convince to be a suitable way as the decision tree is not flexible algorithm for possibility prediction problems.

3. Symptom-Disease Knowledge Graph Creation Framework

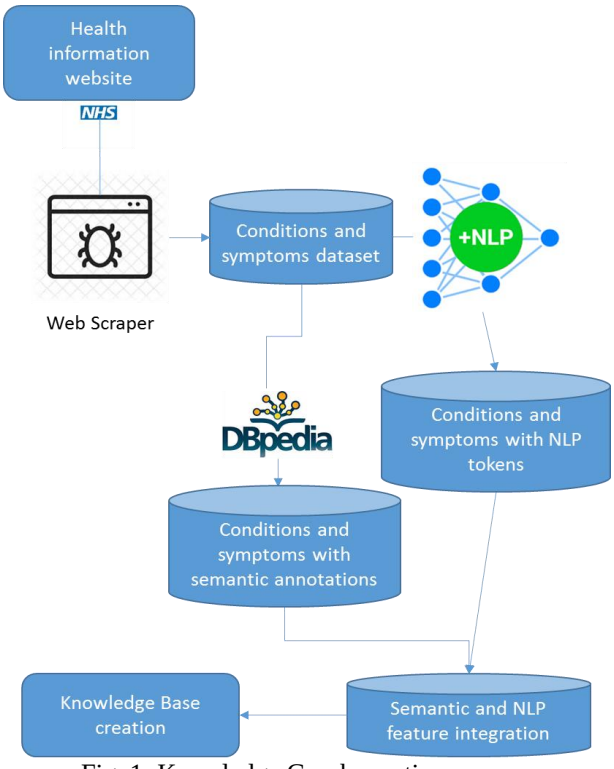

Fig. 1: Knowledge Graph creation process

The feature enrichment engineering process on the symptom description include four major steps (see Figure 1):

- Crawling data from trust source to generate symptom and condition relational dataset.
- Applying NLP word token and sentence token abstractions to highlight meaningful words and phrases in the symptom description (Figure 2).

- Obtaining semantic annotations that related to the condition and symptom descriptions from DBpedia knowledge graph via DBpedia spotlight API [6] (Table 1).
- Finally the NLP tokens and DBpedia annotations are integrated to a semantic framework that can be queried using SPARQL and applied to ML approaches.

	1	2	word token
1	Abdominal aortic aneurysm	a pulsating feeling in your stomach (abdomen),...	[a, pulsating, feeling, in, your, stomach, abd...
2	Acne	periods – some women have a flare-up of acne j...	[periods, some, women, have, a, flare, up, of,...
3	Acute cholecystitis	a high temperature (fever),nausea and vomiting...	[a, high, temperature, fever, nausea, and, vom...
4	Acute lymphoblastic leukaemia	pale skin,feeling tired and breathless,having ...	[pale, skin, feeling, tired, and, breathless, ...
5	Acute lymphoblastic leukaemia: Children	red blood cells, which carry oxygen around the...	[red, blood, cells, which, carry, oxygen, arou...
6	Acute lymphoblastic leukaemia: Teenagers and y...	white blood cells – which help us fight infect...	[white, blood, cells, which, help, us, fight, ...
7	Acute myeloid leukaemia	pale skin,tiredness,breathlessness,a high temp...	[pale, skin, tiredness, breathlessness, a, hig...
8	Acute myeloid leukaemia: Children	red blood cells, which carry oxygen around the...	[red, blood, cells, which, carry, oxygen, arou...
9	Acute myeloid leukaemia: Teenagers and young a...	white blood cells – which help us fight infect...	[white, blood, cells, which, help, us, fight, ...
10	Acute pancreatitis	nausea (feeling sick) or vomiting,diarrhoea,in...	[nausea, feeling, sick, or, vomiting, diarrhoe...
11	Addison's disease	fatigue (lack of energy or motivation),letharg...	[fatigue, lack, of, energy, or, motivation, le...
12	Alcohol misuse	breathing,heart rate,gag reflex, which prevent...	[breathing, heart, rate, gag, reflex, which, p...
13	Alcohol poisoning	confusion,severely slurred speech,loss of co-o...	[confusion, severely, slurred, speech, loss, o...
14	Alcohol-related liver disease	abdominal (tummy) pain,loss of appetite,fatigu...	[abdominal, tummy, pain, loss, of, appetite, f...
15	Allergies	swollen lips, tongue, eyes or face,dry, red an...	[swollen, lips, tongue, eyes, or, face, dry, r...

Fig. 2: Symptom-disease dataset with NLP analysis

Table 1 Semantic annotation example

Common Cold	Word token semantic annotation	DBpedia linked abstract semantic annotation
Semantic enrichment	http://dbpedia.org/resource/Rhinorrhea http://dbpedia.org/resource/Sneeze http://dbpedia.org/resource/Dysphonia http://dbpedia.org/resource/Thermoregulation http://dbpedia.org/resource/Fever http://dbpedia.org/resource/Myalgia http://dbpedia.org/resource/Blood_pressure http://dbpedia.org/resource/Fever	http://dbpedia.org/resource/Viral_disease http://dbpedia.org/resource/Infection http://dbpedia.org/resource/Respiratory_tract http://dbpedia.org/resource/Paranasal_sinuses http://dbpedia.org/resource/Larynx http://dbpedia.org/resource/Sore_throat http://dbpedia.org/resource/Rhinorrhea http://dbpedia.org/resource/Sneeze http://dbpedia.org/resource/Headache http://dbpedia.org/resource/Fever … (more)

4. Experimental of data exploration on the dataset

The symptom descriptions for each health condition gathered from data crawling step are purely text for human, which will cause difficulties to ML algorithms to do learning tasks because of noisy words. For example, without NLP processing, the five most common words of appeared in symptoms are “the” (1317 times), “and” (977 times), “of” (955 times), “or” (901 times) and “a” (901 times).

Therefore, the word tokens are produced as the first try to lift the meaningful words that indicate the symptoms. With NLP, the most common words in symptom descriptions are “pain” (199 times), “feeling” (184 times), “skin” (178 times), “may” (171 times), and “loss” (161 times). However, having single word tokens for the condition sometimes looses certain relations between verb and nuns. For example, “feeling pain” makes more sense than separating them into two independent words. The other word tokens are generated in order to get phrases if single word token is not meaningful. With optimizing

of word tokens and sentence tokens, the most common 200 symptoms are displayed in Figure 4 word cloud. Clearly, the "loss of appetite", "diarrhoea", "tiredness", "feeling sick", "weight loss" and "dizziness" as well as "high temperature" are the most appeared symptoms for conditions, which matches our exceptions. The same experiment has been applied to semantic annotation as well. Figure 3 shows clearly the major key words or semantic terms from overall dataset.

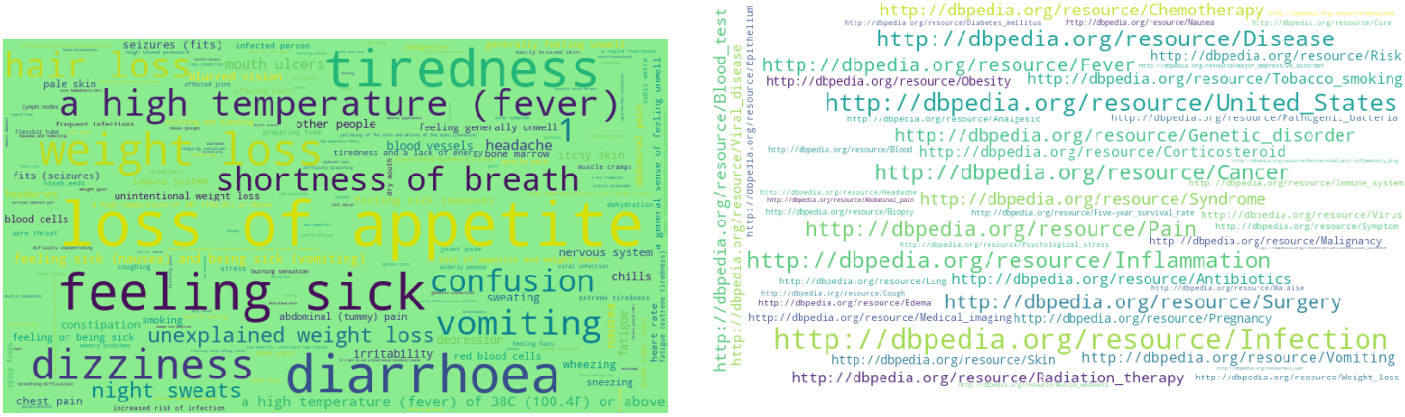

Fig. 3 WordCloud for NLP word token and DBpedia semantic terms inside the tripe dataset

5. Conclusion

In this paper, we present a symptom-disease data mining framework to enable generating a semantic linked knowledge graph for common health conditions. At the moment we generated a 22431 triple links with relations between symptom and disease or disease to disease. The next step research will on the further enrichment on the triple dataset to create more meaningful relations to support ML prediction and clustering.

References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955

[2] N. Papachristou et al., "Comparing machine learning clustering with latent class analysis on cancer symptoms' data," 2016 IEEE Healthcare Innovation Point-Of-Care Technologies Conference (HI-POCT), Cancun, 2016, pp. 162-166.

[3] Harini D. K., Natesh M., Prediction of Probability of Disease based on Symptoms Using Machine Learning Algorithm, International Research Journal of Engineering and Technology(IRJET), e-ISSN: 2395-0056, Volume: 05, Issue: 05, May-2018.

[4] Md. Tahmid Rahman Laskar, Md. Tahmid Hossain, Abu Raihan Mostofa Kamal and Nafiul Rashid. Article: Automated Disease Prediction System (ADPS): A User Input-based Reliable Architecture for Disease Prediction. International Journal of Computer Applications 133(15):24-29, January 2016.

[5] Shratik J. Mishra, Albar M. Vasi, Vinay S. Menon ,Prof. K. Jayamalini, GDPS -General Disease Prediction System, International Research Journal of Engineering and Technology(IRJET), e-ISSN: 2395-0056, Volume: 05, Issue: 03, May-2018.

[6] https://www.dbpedia-spotlight.org/

[i] https://www.kaggle.com/