Extracting and Representing Causal Knowledge of Health Conditions

Hong Qing Yu

University of Bedfordshire, School of Computer Science and Technology, Luton, UK
hongqing.yu@beds.ac.uk

Abstract. Most healthcare and health research organizations published their health knowledge on the web through HTML or semantic presentations nowadays e.g. UK National Health Service website. Especially, the HTML contents contain valuable information about the individual health condition and graph knowledge presents the semantics of words in the contents. This paper focuses on combining these two for extracting causality knowledge. Understanding causality relations is one of the crucial tasks to support building an Artificial Intelligent (AI) enabled healthcare system. Unlike other raw data sources used by AI processes, the causality semantic dataset is generated in this paper, which is believed to be more efficient and transparent for supporting AI tasks. Currently, neural network-based deep learning processes found themselves in a hard position to explain the prediction outputs, which is majorly because of lacking knowledge-based probability analysis. Dynamic probability analysis based on causality modeling is a new research area that not only can model the knowledge in a machine-understandable way but also can create causal probability relations inside the knowledge. To achieve this, a causal probability generation framework is proposed in this paper that extends the current Description Logic (DL), applies semantic Natural Language Processing (NLP) approach, and calculates runtime causal probabilities according to the given input conditions. The framework can be easily implemented using existing programming standards. The experimental evaluations extract 383 common disease conditions from the UK NHS (the National Health Service) and enable automatically linked 418 condition terms from the DBpedia dataset.

Keywords: Knowledge Graph · Causality · Health · NLP · AI

1 Introduction

There are many high-quality health condition data available online, such as the UK website of National Health Service and condition descriptions on Wikipedia. Understanding the causal relations inside this data will be useful to enhance self-healthcare awareness and education. The research problem is how to extract these causal relations automatically and understand the semantics from this data e.g. sentences and paragraphs. For example, extracting Pneumonia is a kind of disease and the coronavirus another kind of disease is one of the causes

to Pneumonia from the sentence "Pneumonia can be caused by a virus, such as a coronavirus (COVID-19)". Besides, the probability is also an important aspect of the causality due to uncertainties e.g. pneumonia can be caused not only coronavirus but also other bacterial infections. In this paper, a probability-based causality extracting and modeling framework is proposed to address this research problem. Two major novelties of the paper are:

(1) A formal health causality extracting framework is proposed to support causal recognition, knowledge modeling, and runtime probability creation.

(2) The first causal knowledge graph is created containing 383 health conditions from the UK NHS website with causal links to 418 Wikipedia health terms through DBpedia annotations.

Rest of the paper has further 4 sections:

Section 2 will discuss some related work.

Section 3 will explain the whole framework and each of the steps.

Section 4 will show the insight evaluation of the generated causal knowledge graph.

Section 5 will present the conclusion.

2 Related Work

Representing health knowledge in a way that the machine can easily process is an important research area. The core topics in the field can be categorized into three major groups.

One is focusing on representing clinical data as knowledge, e.g. Electronic Health Records (EHR). An integration process to build a common data model was proposed by [1, 2] aimed to produce shareable, transportable, and computable clinical data. However, the work only emphasized the system architecture level (NoSQL) and data representation level (RDF) but did not directly address knowledge understanding especially the causal relations. Many different frameworks worked in this direction of ontology development and triple populating.

The second category is to apply state of art machine learning approaches to the existing KG data to perform prediction or classification tasks. The paper [3] proposed a medical code prediction framework to build a KG with NLP and external Wikipedia semantic links to the information source. The prediction results through graph vector encodings applied to the logistic regression classification algorithm. However, these knowledge prediction approaches lack of explanations and tractability. Moreover, they still can not tell the causes of such a prediction.

The last direction is to directly add causal knowledge to the data. This type of research can be traced back to the 1980s as Neyman-Rubin causal inference theory was published. However, the concepts of causal and association or correlation are always been mixed or misunderstood until the formal mathematics models are represented by Pearl in [4]. The model computes probability joint distribution on the directional graph that satisfies the back-door criterion, which is a do(X=x) rather than a random x to have a probability prediction on Y

based on statistic knowledge. In simplified terms, the causal relation should be observed if one property were modified, then the other property of a probability distribution would also change. Therefore, we can distinct associational relations and causal relations. Most recently, this idea has been applied on top of the reinforcement learning process by the DeepMind team [5]. At the same time, some work starts to investigate an approach to add probability concepts into knowledge graphs to express knowledge with belief rating thresholds. Based on this idea, a Probabilistic Description Logic (PDL) was explained in 2017 [8] to deal with subjective uncertainty. The PDL extended Tbox and Abox definitions in the classic Description Logic (DL) with probabilistic thresholds notations. However, the probability needs to be defined at design time or from current knowledge not able to be tuned dynamically. In addition, it completely does not model the causal relations but is replaced by probability.

3 Causality Knowledge Extracting and Modelling

Overall, the causality extracting framework contains four major approaches as shown in Fig. 1. The CNN algorithm is applied to identify the sentences that contain causal relations. The composition of the NLP and semantic annotation process is developed in generating semantic word tokens. The causality description logic is introduced to guide the causality knowledge graph generation by lifting the semantic word tokens. Finally, the runtime probability knowledge graph with defined probabilities will be created when certain inputs values are calculated accordingly.

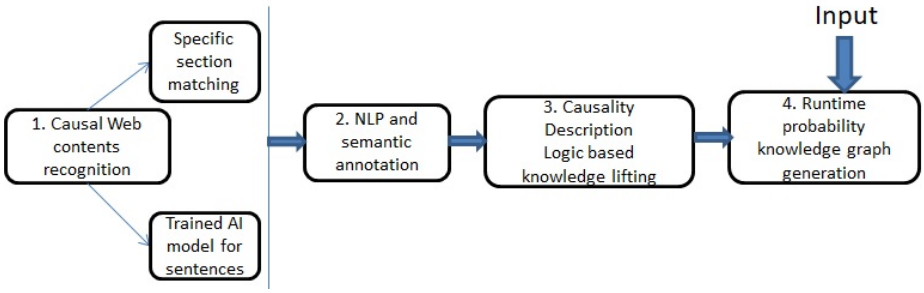

Fig. 1. Four approaches of the framework

3.1 Causality recognition

Two methods have been applied in this approach. One is to directly believe that certain sections of the web contents that should contain causality knowledge. For example, the symptom and causes sections, which can be defined based on the

research of interests. The other method is to build a recognition AI model that can identify sentences that has causality statement(s). There are two most recent research results shows using self-attention deep neural networks can achieve more than 70 percentage accuracy on this task [6, 7]. However, the scenarios are more complex to detect and categorise multiple causal effect classes. In addition, these algorithms are too expensive in terms of computing resources and time. Our task majorly tells if the sentence contains causality that is a binary question. A cheap solution is also the requirement in our scenario. To achieve it, five different machine learning algorithms have been evaluated based on a training dataset. The training dataset is composed of two datasets from the previous research work presented in [9]. Table 1 shows that CNN model provided the best result of recognising causal sentences.

Table 1. AI algorithms evaluation

Algorithms	Total accuracy	F1 score	CPU/GPU	library
Random Forest	0.79	0.79	CPU	Scikit-learn
SVM	0.81	0.81	CPU	Scikit-learn
Logistic Regression	0.81	0.81	CPU	Scikit-learn
MNB	0.81	0.81	CPU	Scikit-learn
LSTM	0.88	0.86	GPU	Keras 2.0
CNN	0.98	0.90	GPU	Keras 2.0

3.2 Causality knowledge modelling

The DL-ιt expression is refined to define Causal Probability Knowledge Base (CPKB) that has four elements as equation 3.2 represent:

$$CPKB = \{T, A, \Phi, P(\phi)\} \tag{1}$$

Where T is the T-box ontology (Terminology structure). A is the A-box instance (Assertions) and Φ is the root causal function that is the major extension to traditional DL-ιt. Φ presents the causal relation that can be happened between any concepts defined inside T. A subclass of Φ can be defined to indicate specific causal relations between two concepts. $P(\phi)$ tells the probability values of causal relations between two instances at Abox level and importantly only at runtime. A set of runtime $P(\phi)$ is calculated based on the input observations.

For the health condition application scenario, Fig. 2 presents the defined T-box and Φ in OWL schema that includes twelve concepts and ten causalities (Φ) and three normal relations.

3.3 Causality extraction and lifting process

The causality extraction process has two components:

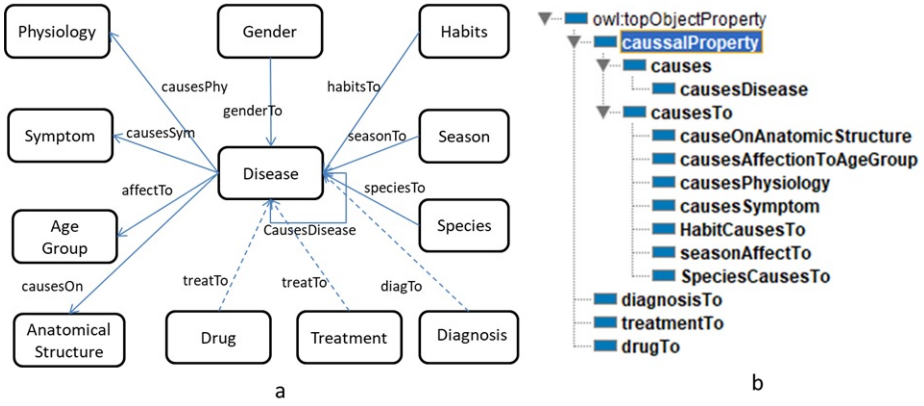

Fig. 2. Health condition CPKB definitions

(1) NLP-based causal keywords tokenization is to capture the keywords that may have causal relations in the identified causality texts from previous steps. The tokenization follows classic NLP steps of segmentation, word tokenization, remove stop words, stemming, and eventually get the noun keywords or phrases. For example, the words of pneumonia, virus, and coronavirus will be captured from the sentence of "Pneumonia can be caused by a virus, such as a coronavirus (COVID-19)"

(2) Semantic lifting calls semantic annotation API (DBpedia spotlight) to classify the keywords and phrases into different terms defined in CPKB ontology based on the RDF: type and other related predictions described in the DBpedia dataset. For example, the word 'Lung' is a type of DBpedia anatomical structure class defined by RDF: type of lung RDF data.

Based on the above two components, we can extract causality for given sentences or paragraphs. In the end, we can generate a knowledge graph for each crawled health conditions from these CPKB based semantic populating. Currently, 383 health conditions' knowledge graph is integrated from the UK NHS website with additional causal semantic links to 418 Wikipedia health terms through the DBpedia dataset.

3.4 Causality-based runtime probability knowledge graph

With health condition causality knowledge in hand, the runtime probability knowledge graph can be dynamically generated based on the numbers of income links to each of the inputs. For example, the input observed conditions for a boy (child and male) are:

Symptoms: cough, breathing, fever, heartbeat, chest pain, fatigue, and shivering, infection. Unwell body position: lung.

With these input conditions, the Fig. 3 (the partial graph of the actual graph as an example) presents a runtime probability distribution among relevant causal

relations. For instance, the Pneumonia disease has around 0.0054 and 0.018 causal probabilities for problems of Heartbeat and Cough respect.

```
<https://www.nhs.uk/conditions/Pleurisy> a <http://dbpedia.org/ontology/Disease> ;
    ns1:hasCausalProbablity [ ns1:causalityTo <http://dbpedia.org/page/Chest_pain> ;
            ns1:pvalue "0.022222222222222223" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Lung> ;
            ns1:pvalue "0.022222222222222223" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Cough> ;
            ns1:pvalue "0.01818181818181818" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Breathing> ;
            ns1:pvalue "0.017857142857142856" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Infection> ;
            ns1:pvalue "0.005376344086021506" ] .

<https://www.nhs.uk/conditions/Pneumonia> a <http://dbpedia.org/ontology/Disease> ;
    ns1:hasCausalProbablity [ ns1:causalityTo <http://dbpedia.org/page/Infection> ;
            ns1:pvalue "0.005376344086021506" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Cough> ;
            ns1:pvalue "0.01818181818181818" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Heartbeat> ;
            ns1:pvalue "0.047619047619047616" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Breathing> ;
            ns1:pvalue "0.017857142857142856" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Chest_pain> ;
            ns1:pvalue "0.022222222222222223" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Fever> ;
            ns1:pvalue "0.008333333333333333" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Lung> ;
            ns1:pvalue "0.022222222222222223" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Unwell> ;
            ns1:pvalue "0.037037037037037035" ] .

<https://www.nhs.uk/conditions/Poisoning> a <http://dbpedia.org/ontology/Disease> ;
    ns1:hasCausalProbablity [ ns1:causalityTo <http://dbpedia.org/page/Child> ;
            ns1:pvalue "0.007518796992481203" ] .

<https://www.nhs.uk/conditions/Polio> a <http://dbpedia.org/ontology/Disease> ;
    ns1:hasCausalProbablity [ ns1:causalityTo <http://dbpedia.org/page/Infection> ;
            ns1:pvalue "0.005376344086021506" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Fever> ;
            ns1:pvalue "0.008333333333333333" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Breathing> ;
            ns1:pvalue "0.017857142857142856" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Childhood> ;
            ns1:pvalue "0.007518796992481203" ] .

<https://www.nhs.uk/conditions/Polymorphic-light-eruption> a <http://dbpedia.org/ont
    ns1:hasCausalProbablity [ ns1:causalityTo <http://dbpedia.org/page/Fever> ;
            ns1:pvalue "0.008333333333333333" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Male> ;
```

Fig. 3. Runtime probability knowledge graph example

4 Insight of Causality Knowledge Graph

After crawled health conditions throughout the NHS webpages and built semantic causal relations with Wikipedia definitions and DBpedia terms, we generated a causality knowledge graph that contains 801 health conditions, 1078

symptoms/physiologies, 377 treatments including drugs, 8 categorized habits, 66 different human groups, and 113 species.

Fig. 4 shows 25 symptoms or physiological reflections that have the most connections with other health conditions. Interestingly, Schizophrenia a kind of mental health condition can be developed from 264 diseases. The other noticeable information is that many diseases may have sequela and contribute to rare diseases. The figure also indicates that diabetes is one of the most common symptoms of other diseases.

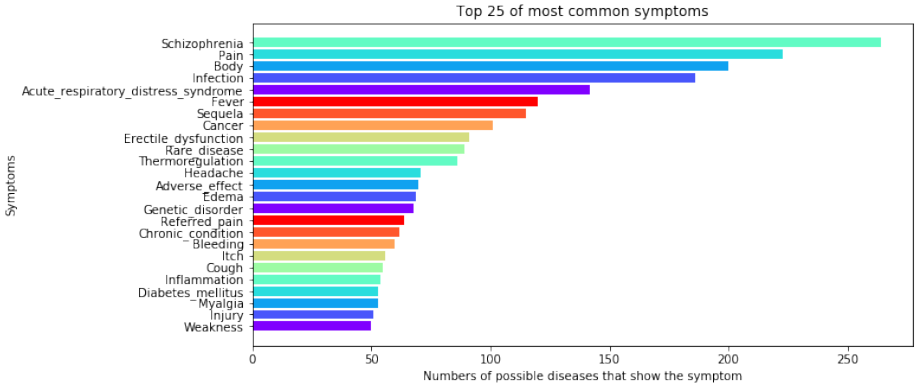

Fig. 4. Top 25 symptoms or physiological reflections

Based on the causal relations, eight habits or lifestyle-related scenarios can contribute to developing serious health problems. The top one is the smoking-related habits are most dangerous and connect to more than 100 diseases. The other noticeable one is overeating.

The causal reasoning result also shows that Autumn and Winter have the most connections to diseases than other seasons which reflects common sense.

Through causal relations, the condition chain is discovered. For example, Rheumatoid arthritis → Psoriasis → Pagets disease nipple → Breast cancer → Weight loss. 3683 5-length-chains, 3847 4-length-chains, and 111186 3-length chains are discovered so far. All these condition chains are the hidden knowledge that is not identified in the original description on the webpages.

Besides, the health conditions from NHS are clustered into 42 groups when applying unsupervised K-mean clustering algorithm and cluster optimization process. For example, a list of observations ['headache, influenza, fever, throat, children'] is mostly related to the health condition in Cluster 0 that contains 12 diseases of ['Bornholm-disease', 'Common-cold', 'Diphtheria', 'Chickenpox', 'Flu', 'Hand-foot-mouth-disease', 'Polio', 'Q-fever', 'Roseola', 'Rubella', 'Slapped-cheek-syndrome', 'Tonsillitis'].

5 Conclusion and Future Work

A causality focused knowledge graph generation approach is introduced in this paper. The major purposes of the work are to extract causal relations inside the health descriptive data on the Web and to create a probability knowledge space at runtime to support further AI tasks. The evaluations on the causal probability knowledge graph have already shown some interesting conclusions and the ability to enhance explanation capabilities of prediction and clustering approaches. The implementation code and the dataset are available at [10]. There are two limitations at current state of art. The first one is that some combination key words e.g. Body pain have not been captured using classic NLP and semantic annotation processes. The second one is that our knowledge has not fully connected to external exist health knowledge datasets e.g. UMLS [11]. In the short-term, our research will focus on addressing these limitations. The long-term future research has a couple of directions. Firstly, to develop an efficient embedding method that can contain causal relation features and apply well-studied machine learning algorithms especially the deep learning architectures. Secondly, to investigate the graph-based learning algorithm that can directly work on the graph data and get utilization from the reasoning power from the graph, causal relations, and runtime probability definitions.

References

1. Overhage, J. M., Ryan, P. B., Reich, C. G., Hartzema, A. G., Stang, P. E.: Validation of a common data model for active safety surveillance research. Journal of the American Medical Informatics Association : JAMIA, 19(1), 54–60, 2012. https://doi.org/10.1136/amiajnl-2011-000376
2. Rosenbloom, ST., Carroll, RJ., Warner, JL., Matheny, ME., Denny, JC.: Representing Knowledge Consistently Across Health Systems. Yearb Med Inform. 2017;26(1):139-147. https://doi.org/10.15265/IY-2017-018
3. Bai, T., Vucetic, S.: Improving Medical Code Prediction from Clinical Text via Incorporating Online Knowledge Sources. In The World Wide Web Conference (WWW '19), Ling Liu and Ryen White (Eds.). ACM, New York, NY, USA, 2019, 72-82. https://doi.org/10.1145/3308558.3313485
4. Pearl, J. 2010. An Introduction to Causal Inference. The International Journal of Biostatistics. 6, 2 (2010).
5. Dasgupta, I., Wang, J., et al: Causal Reasoning from Meta-reinforcement Learning. arXiv preprint 2019, arXiv:1901.08162.
6. Li, Z., Li, Q., Zou, X., Ren, J.: Causality Extraction based on Self-Attentive BiLSTM-CRF with Transferred Embeddings, 2019 arXiv:abs/1904.07629.
7. Dasgupta, T., Saha, R., Dey, L., Naskar, A.: Automatic Extraction of Causal Relations from Text using Linguistically Informed Deep Neural Networks, 2018, 306-316. 10.18653/v1/W18-5035.
8. Gutierrez-Basulto, V., Jung, J.C. and Lutz, C.: Probabilistic Description Logics for Subjective Uncertainty. Journal of Artifical Intelligence Research 58, 2017, 1-66.
9. Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., Weischedel, R.: Ontonotes: the 90 percentage solution. In Proceedings of NAACL, Companion Volume: Short Papers, 2006, pages 57-60. ACL.

10. NHS causal knowledge graph with evaluation and clustering, https://github.com/semanticmachinelearning/nhscausalknolwedgegraph. Last accessed 21 June 2020
11. Olivier, B.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32(Database issue):D267-D270.