Accelerating Epidemiological Investigation Analysis by Using NLP and Knowledge Reasoning: A Case Study on COVID-19.

abstract:

Triple-based Multi-Task Neural Network (TMT-NN)
jointly recognizing epidemiological entities and relations from case reports.
an epidemical knowledge graph and the corresponding inference engine are built to recover the infection modes, sources and pathways.
OpenKG.cn
Covid-19
travel trajectories before the disease.
including demographic data & clinical manifestations of confirmed cases, travel trajectories before the disease.
人口统计数据 确诊病例的临床表现 确诊前发生的活动轨迹数据
tracing and analyzing infection source and pathway for each patients, then discovering & generating epidemiological insights with visualization.
追踪 分析 感染源 路径 发现与可视化.
related work:
① current methods for epidemiological investigation analysis.
② state of art for medical information extraction.
③ applying knowledge graph in epidemiology research
IE: automatic extraction of concepts, entities, events, as well as relations & associated attributes from free texts.
general work in this paper.
① design & implement a neural network based on BERT.
② an inference engine is constructed to discover infection sources, pathways and networks based on KG.
推理引擎

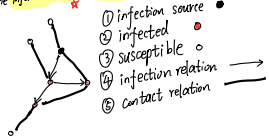
System Description & epidemiological KG:
an epidemiological case report for an individual patient usually includes 4 types of information:
① demographics 人口统计数据
② recent native log 近期活动记录
③ disease related onset, diagnosis and treatment information 新冠疾病的症状,诊断和治疗信息
④ the information of close contacts 密切接触者的信息

Epidemiological information extraction via a novel deep neural network specifically designed for case reports.
nested tuples.
an epidemiological case reports can be structured
{ 1 patient info tuple: (patient ID, age, gender, residence place, epidemiological contact history).
1/more event tuple: (time, location, verb, vehicle, manifestation, persons)
1/more social relation tuples: (id, name, relation-type) }

TMT-NN is specifically designed with a hierarchical entity labelling structure to extract corresponding nested tuples and their corresponding attributes in an end to end way.
分层实体标注结构
encoding → tuple keyword labeling → position embedding in Transform → tuple attribute labeling.

Infection Network Inference & Infection Source Identification.
感染网络推断 感染源识别.
structured epidemiological information are organized as knowledge graph, we use epidemiological inference algorithm to discover the implicit infection relations among confirmed patients and find out all possible infection sources.

Problem Description.
$V = \{1, \dots, N\}$: N individual patients.
$G_{infect} = (V, E_{infect})$: directed graph
$E_{ij} = [i,j]$: directed edge, infection relation.
$G_{contact} = (V, E)$: undirected graph
$E_{ij} = (i,j)$: undirected edge, contact relation .... with a independent probability p.
Contact: the occurrence of a physical association of two individuals that could be sufficient for disease transmission.
though not all contacts between infective & susceptible individuals are guaranteed to result in transmission.
★ the infection network is a subgraph of the contact network, shown in Figure 1.
epidemiological investigation, case reports ⟹ infection network from contact network. ⇒
Our goal is to infer the infection source & transmission paths (pathway along which the epidemic spreads) from the contact network with the history observation $O(t)$.
the inferred source(s), transmission paths with nodes are the directed subgraph of the undirected contact network. we call it directed infection network (G_{infect} or g ★



$$\hat{s}, \hat{g} = \arg\max_{s \in V, g \subset G_{contact}} P(O|s,g)$$

s: infection source
g: inferred infection network
$P(O|s,g)$: possibility of observation O of all nodes states after the disease transmits in contact network.

4 steps pipeline:
① construct the contact network from structured case reports stored in KG. [with the state of each node at every points of time t_i]
* the individuals in the epidemiological KG are the nodes in the contact network.
contact relation ① social relationship between two patients.
② the contact events associated two or more patients who have least interacting event(s) with the same time duration & location.
social relationship / contact event : undirected edge
Notes: ① may not be connected graph.
② may exist loops.
2* spot nodes. the shared spots in the same time t or the same duration of events can be a transmitting mediator of the disease.
add spot nodes to the node set V.

E_{x2} E_{y2}

3* transmission time t_{ij} for each node. After this time, the node state changes from susceptible to infected.
Note: since the source node is not defined yet, we use absolute time instead of relative time from source node.
individual node: onset date of reported case as the time of each node.
spot node: time of associated event.

② Divide the contact network into connected subgraphs to reduce computation complexity.

③ For each subgraph, identify the possible infection sources and transmission paths to get infection subgraph.
⇒ identify all combined alternatives of s & g
⇒ $\hat{s}, \hat{g}$ which maximize $P(O|s,g)$

E_{xy}: | x→y
∅ no infection relation
↓ y→x

Rule1: For any $x \in V$, set $P(O|s) = p$ [p: avg. possibility value of all nodes].
if x has contact history with confirmed case(s) before time t, set $P(O|s) = w_1 p$.
if x has contact history with suspected case(s) or person(s) come from epidemic area(s) in [two time t, where $w_1 > w_2 > 1$

if x has contact history with suspected case(s) or person(s) come from epidemic area(s) or been to epidemic area(s) before

w_1, w_2 are the empirical constant value.

Rule2: For any $x, y \in V$, contant edge $E_{xy} \in E$, t_x: onset time of x, t_y: onset time of y.

L: The directed infection edge L_{xy}: $x \to y$

L_{yx}: $y \to x$

$P(O|L_{xy}) > P(O|L_{yx})$ if $t_x < t_y$.

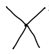
Rule3: For any $x, y, z \in V$, $E_{xy} \in E$, t_x: onset time of x, t_y: onset time of y

if $t_x = t_y$ & $P(O|x) > P(O|y)$, $P(O|L_{xz}) > P(O|L_{yz})$

Rule4: For any $x, y, z \in V$, $E_{xz}, E_{yz} \in E$, the onset time t_x, t_y, t_z.

if $t_x < t_y < t_z$, $P(O|L_{xz}) > P(O|L_{yz})$

$$P(O|s,g) = \sum_{E_{xy} \in E} P(O|L_{xy})$$



注：1个点当有多个入边时 只有概率最大的边可以被加入计算 [一个病人只能被感染一次].

$\underset{L,V}{\arg\max}\ P(O|s,g)$ L: set of directed edge, V: nodeset } directed infection subgraph. $\Longrightarrow$ source [the node which in-degree=0].

④ Validate the inferred graphs. ---- 验证

if conficts happens, we should check the result. [conflictions between original epidemiological investigation data & inference result.] [maybe caused by wrong reported date, etc]. $\Downarrow$ manually to fix

Dong Liu

2021.11.26