面向英文新冠推文的实体标注规范[1]

YLab
浙江大学
版本： 1.0

[1]本规范的制定参考自 Zhang H, Zong Y, Chang B, et al. 面向医学文本处理的医学实体标注规范 （Medical Entity Annotation Standard for Medical Text Processing）[C]//CCL. 2020: 561-571.

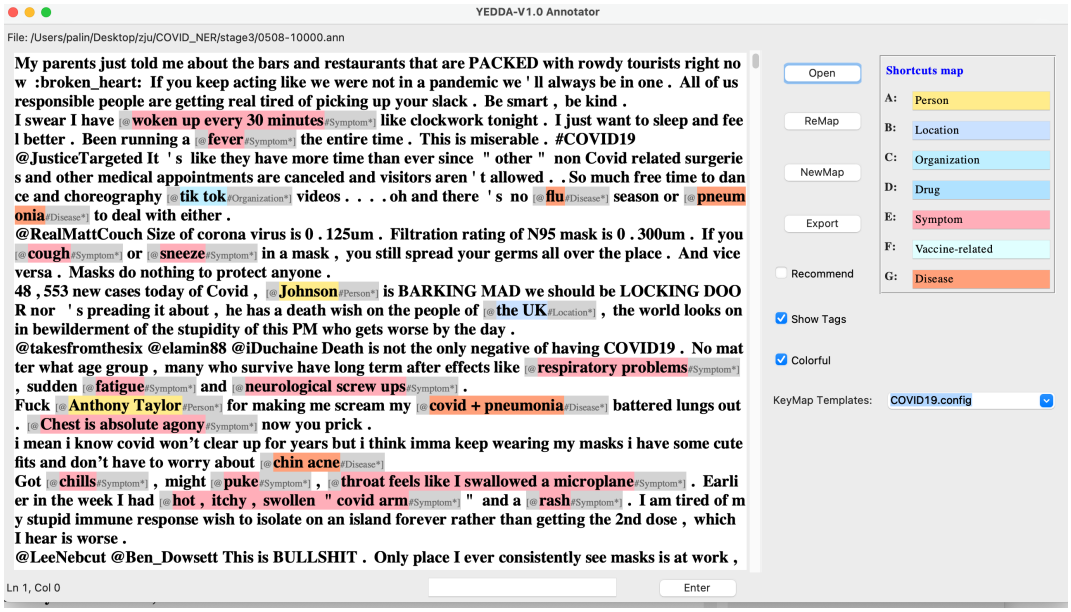
目录

背景介绍

近几年来，新型冠状病毒肺炎（COVID-19）在全球不断蔓延，由此引发人们在社交媒体上对新冠肺炎相关话题的广泛讨论。为了研究 COVID-19 对人民生活的影响，如何有效地利用命名实体识别（NER）技术来分析公众对社交媒体上与新冠相关的实体（如药物、疫苗）的关注程度和态度就显得至关重要。然而，现有命名实体识别（NER）模型所使用的训练数据大多集中在通用领域，只支持诸如人名、地名等通用实体类型的提取，无法从文本中识别出和医学相关的实体信息，极大地限制了 NER 模型在医学领域的应用。此外，社交媒体文本的表达形式自由多变，并未严格遵循语法规则，这进一步加大了 NER 的识别难度。针对以上问题，本指南提出从医学研究角度出发，帮助研究人员构建一个面向新冠相关推文的 NER 数据集，用以开发更好的医学社交媒体文本理解工具，推动包括流行病学研究在内的计算社会科学研究的发展。

标注流程

如下图所示，本项目的所有标注工作均使用 YEDDA 标注平台（https://github.com/jiesutd/YEDDA）完成。



本指南将标注过程划分为以下三个阶段：

1. 在预标注阶段，要求所有标注人员进行 3 轮标注。标注一致性使用 F1 值来度量。F1 大于 80%的标注人员将被选中进入正式标注过程。在整个过程中，标注指南也会不断更新。
2. 在正式标注阶段，将标注人员分为 3 组，每组 2 人，以确保每条推文都被标注两次。当标注出现不一致时，会安排组外标注人员介入以确定推特的最终标注结果。
3. 正式标注阶段结束后，项目组对标注结果进行质量控制检查，确保标注的推文符合标注指南的要求。最终标注人员之间的一致性为 85.0%。

实体类型描述

实体类型	标签	备注	涵盖范畴	举例
人名	PER	Person	真实人物、虚构人物、文学作品中的人物以及宗教人物（如God）。	Virginia Wade, Harry Potter
组织机构名	ORG	Organization	公司、学校、医院等。	Google
地名	LOC	Location	国家、省、市等。	France, China
症状	SYM	Symptom	临床表现，泛指患者不适感觉以及通过检查获知的异常表现。	Fever，headache
药物	DRU	Drug	用来预防、治疗及诊断疾病的物质，包括临床药物、抗生素等。	Dexamethasone
疫苗相关	VAC	Vaccine-related	包括疫苗类型和疫苗品牌的描述。	Moderna
疾病	DIS	Disease	导致病人处于非健康状态的原因或者医生对病人做出的诊断,并且是能够被治疗的，包括疾病或者综合征（（disease or syndrome ）、中毒或受伤（injury or poisoning ）等。	cardiopathy ，hypertension

实体标注准则

实体标注通则

1. 应该是**具体的、特定的**，而不是抽象的、泛指的。
   比如：”woman, girl, place, location, father, small town，hospital、school、college”等就不应标注。
2. 若出现标注实体的**英文缩写或者俗称**，均需要标注。
3. **不允许**实体嵌套，换句话讲，只标一个实体的最长边界，不标注内部包含的其他实体。
4. 标注时不用将多余的空格也标进来，即标注的单词范围头尾不应出现空格。
5. 标注时一定要**把输入法切换为英文**，防止标注时词语被错误替换。
6. 标注时**以完整单词为单位**，不应截取单词的部分标注，即一个词要么全部属于一个实体，要么全部不属于。
7. 当开启 YEDDA 的推荐模式时，对于推荐的标注均需进行**人工确认**，否则最终的标记文件会自动记录所有推荐的结果，从而影响标注的准确性。
8. 只标注 Location 和 Organization 中出现的 the。其他实体类型中的 the 一律不用标注。
   比如：the U.S., the U.K., the WHO、the NIH、the Westchester Country Center
9. 修饰症状严重程度的词语**无需**标注，只标注症状本身，但是出现身体部位时应标注。

实体标注细则

1. 人名（PER）

人名包括真实人物、虚构人物、文学作品中的人物以及宗教人物的名字。如果没有出现具体的名字，则不要标注。

- 一些容易漏标/错标的示例

标注示例	**说明**
The Queen has said...	The Queen 不是具体人名
Mr [@**Weir**#Person*] also told	称谓不需要标注
allergist [@**Troy Baker**#Person*], MD, tells what to expect and explains that serious allergic reactions are extremely rare	只标注人名，不要标注身份
[@**GOP**#Organization*], MAGA leaders，Qleaders said	未出现人名，故无需标注
[@**Katten**#Organization*] lawyers offer [@**UK**#Location*] employers as the best approaches towards COVID19 vaccination in the	没出现具体名字时不用标注，比如 katten 律所的律师，并未

workplace.	出现具体的人名，故只需标注组织即可，同理“英国雇主”中也没有出现具体的人名，因此只需标注“英国”这一地名。
How a secret military experiment left **Black Georgians** wary of COVID-19 vaccines	Black Georgians 表示人种，并未出现具体人名，故不用标注
now **god** knows who he is spread it to	上帝不需要标
Excellent news for your sister [@**Teresa**#Person*]!	
Congratulations, [@**Graydon**#Person*]!	
I am afraid I may have mutated to some dark reality mutant [@**Ashley**#Person*].	
[@**Lil**#Person*] side effects is not much compared to lasting effects of covid	
This mother [@**Miriam**#Person*]" is a schismatic who ignores the teaching of the Church,	只标注人名，不标注角色

2. 地名（LOC）

地名包括洲、海洋、国家、省、市、县、地区、街道、乡、镇、村、机场、军事基地、军区、铁路、公路、桥梁、海峡、海湾、港湾、河流、湖、公园、草原、煤矿、牧场、养殖场、音乐厅、剧院、教堂、寺庙、图书馆、博物馆、美术馆、展览中心、公园、动物园、植物园、火车站、广场、大厦、大楼、体育场（馆）、游泳馆（池）、赛车场、商城、超市、书店（城）等城市公共设施，还包括某些特定的城市建筑和虚构的处所。

- 一些容易漏标、错标的例子：

标注示例	**说明**
joining [@**Chicago city**#Location*]... many [@**african countries**#Location*]...	地点+city/province/country 时，应将 city/province/country 一起标注
All these vaccines made by [@**the West**#Location*], didnt harm you.	西方世界
I know [@**BC**#Location*] COVID updates/reporter questions are minimal,	不列颠哥伦比亚（省名，位于加拿大西部）
[@**Mark Drakeford**#Person*] gives coronavirus update as [@**Wales**#Location*] reaches first major vaccine milestone.	
will be crucial in ensuring it reaches all corners of [@**Australian society**#Location*]	
[@**the U.S.**#Location*]	定冠词一起标注
This is the start of the social credit system that [@**China**#Location*] use.	

[@**CDC#**Organization*] guidelines are not followed in [@**Providence#**Location*], [@**RI#**Location*] schools no mandatory testing no vaccines. The risks are is real. Sad!	
[@**Khalistan#**Location*] takes a backseat as [@**Canada#**Location*] approaches [@**India#**Location*] for COVID Vaccines.	

3. 组织机构名（ORG）

An organization is an entity – such as a **company**, an **institution**, or an **association** – comprising one or more people and having a particular purpose.
机构名包括股票（证券）交易所、国家或国际立法部门或行政部门、商业团体（公司、企业、工厂）、电视台、广播电台、报刊杂志、出版社、政党或党派、学校、科研院所、医院、诊所、邮电局、乐队、体育运动队、联盟、议会或代表大会、军队、咖啡厅、酒吧、饭店、旅馆，以及虚构的机构等。

- 对于不确定的缩写，建议使用搜索引擎确定，以免漏标。

标注示例	说明
a signature of its [@**Wuhan lab**#Organization*] creators?	Wuhan lab 应该一起标注为 Organization。
We call on the Mayor, [@TFL#Organization*] and [@Unite the Union#Organization*] to lobby [@**the UK government**#Organization*] with the	国家名+government 应该整体标注为 organization
[@**TNXP**#Organization*] connection to [@Alzheimer#Disease*]'s is pretty tenuous.	**TNXP 是**一家制药公司的缩写
Pains me to hear of [@**NHS**#Organization*] and care staff who refuse the Covid vaccine.	英国国民健康保险制度（National Health Service）
and a [@**Food and Drug Administration**#Organization*] version would not be available for more than a decade	

- 其他容易错标、漏标的例子：

标注示例	说明
a volunteer administering at [@**Health Centre**#Organization*] shares	一般来说组织应该是具体的，比如 xx 医院，xx 军队，只出现医院这个单词时不应该标注。但是如果单词首字母是大写的，比如 Health Centre，则认为其具体指代某个机构，并将其标注为组织。
Calling **811** or the using [@AHS#Organization*]	811 或者 911 不应视作组织
[@**Glasgow Live**# Organization *]	
How [@**retail pharmacies**#Organization*] plan to handle leftover COVID-19 vaccine doses [@**Retail pharmacies**#Organization*] around [@the U.S.#Location*] are taking different approaches to making sure extra COVID-19 vaccine doses do not go to waste, including using waiting lists or giving leftover .	
[@Scotland#Location*] covid update **RECAP**	RECAP 不是 organization

[@**DWP**#Organization*] and other depts can get access to that record.	Department for Work and Pensions （英国）就业和退休保障部）
For the most advanced Covid screening and quarantine management solution, sign up at [@**FeverIQ.com**#Organization*]	
[@**CDC**#Organization*] guidelines are not followed in [@**Providence**#Location*], [@**RI**#Location*] schools no mandatory testing no vaccines. The risks are is real. Sad!	疾病防治中心（Centers for Disease Control）
partner at on a **vaccination rollout guide** to help pharmacies navigate the various state approaches to efforts. Read more	疫苗接种指南不标注
With my co-morbidities and my work in **healthcare**,	不标注
[@**RI doctor**#Person*] reaches out to the [@**Black community**#Organization*] about the COVID vaccine via	
[@**India**#Location*] [@**Dispatch**#Organization*] Long-term [@**loss of smell**#Symptom*] in many, 6 [@**Mumbai**#Location*] wards reported no death last week, Covid allocation can cover 500 million people and more. brings us news relevant to [@**India**#Location*]'s fight against	
Covid Epi Weekly: Best of Times, Worst of Times The third US surge is fading fast but variants, some ominous, are spreading fast. Vaccination is picking up steam but we are failing to address equity and pandemic **fatigue** is high. We must hang on until most of us are vaccinated. 1/	不应该标注
The Holocough - Health care worker dies after second shot of "COVID-19" vaccine via	不应该标注
Old Wine in New Bottles : Low-Tech Approaches to a Covid-19 Vaccine	不应该标注
I heard **[@Florida governor#Person*]** was selling vaccinations to his buddies is why I asked. My wife is fully vaccinated. **[@Madura#Location*]**.	

4. 症状（SYM）

症状是指病人的不适或痛苦表现。通常是病人主观感觉的不适，如腹痛、头晕等，或是自己发现的病理改变，如血尿便血、活动障碍等。

附 1：各种疾病的常见症状包括但不限于*症状列表第四版.csv*【见附件】：
附 2：新冠的常见症状包括但不限于以下列表：

COVID symptoms	
fever	muscle aches
dry cough	muscle pain
tiredness	body aches
loss of taste	congestion
loss of smell	runny nose
aches	nausea
pains	vomiting
headache	diarrhea
sore throat	myalgia
nasal congestion	sputum
red eyes	shortness of breath
diarrhoea	abdominal pain
skin rash	pneumonia
fever	dyspnea
chills	hypoxia
cough	respiratory failure
shortness of breath	shock
difficulty breathing	multiorgan dysfunction
breathing difficulties	fatigue

注：以上附表仅供参考，所有标注者认为是症状的字段都需要标注出来。

- 当症状前后出现身体部位时，要一并标注。表示程度的形容词、副词不用标注。

示例	示例
I was feeling a lot of [@**body ache**#Symptom*] and [@**soreness all over**#Symptom*]这个里面 all over 需要标注	my arm is no longer [@**sore at the site**#Symptom*].
Cue splitting [@**headache**#Symptom*], [@nausea#Symptom*]and the [@shivers#Symptom*]	Some mild [@**headache**#Symptom*]
The second COVID vaccine is giving me	Some mild [@**cough**#Symptom*]

horrible [@**body chills**#Symptom*]	
…my [@**arm was paining**#Symptom*]…	[@**sore right arm**#Symptom*]
I feel normal just a little [@**sore on my arm**#Symptom*] but I am good now	mild [@**body aches**#Symptom*]
My [@**arm was sore**#Symptom*] for 2 days,	slight [@**fatigue**#Symptom*]
Some mild [@**fever**#Symptom*]	slight [@**sore throat**#Symptom*]
low grade [@**fever**#Symptom*]	[@**Fever**#Symptom*] has broke but have terrible [@**head and muscle ache**#Symptom*].
[@**Pain swelling on the arm**#Symptom*] where you got the shot 2. [@**Fever**#Symptom*] or [@**chills**#Symptom*] 3. [@**Tiredness**#Symptom*] 4. [@**Headache**#Symptom*] Remember to get your 2nd shot.	My father in law has been in agonizing [@**pain**#Symptom*] for the last 24 hours
[@**runny nose**#Symptom*]	The 2nd dose of the COVID vaccine must have woken all the dormant peach **takka** and **Smirnoff** in my immune system from 2015 because I [@**woke up feeling drunk** #Symptom*] asf. I was prepared for **a** [@**fever**#Symptom*] and [@**headache**#Symptom*]. Absolutely no one told me to be prepared for senior yr of highschool. Wtf is this
[@**Swollen lymph nodes Lumps under arm or on neck**#Symptom*] Extreme [@**Fatigue**#Symptom*]	Am suffering a lot of [@**pain in my left eyes**#Symptom*] and [@**headaches**#Symptom*] since yesterday as a result of taking the Corona virus vaccine.
[@**Headache**#Symptom*], extreme [@**body ache**s#Symptom*] even my [@**skin hurt**#Symptom*].	True: Earth spiked a bit of a **fever** in 2020, partly because of cleaner air
my [@**left arm still hurts**#Symptom*], like I have been beaten with a bat.	My 2nd dose of the covid vaccine is treating me well! Only having [@**arm soreness**#Symptom*] and feeling [@**fatigued**#Symptom*]/pretty [@**sleepy**#Symptom*].

- 当症状为复数、过去时、进行时，应该将完整词语标注。

示例
…cover [@**coughs**#Symptom*] and [@**sneezes**#Symptom*]…
some just had [@**sore arms**#Symptom*] some just
[@**body aches**#Symptom*]
[@**headaches**#Symptom*]
[@**fevers**#Symptom*]
[@**coughing**#Symptom*]

- 症状之间出现连接词，如 with，and 等，应将症状单独标注。

I feel [@**achey**#Symptom*] with [@**chills**#Symptom*] but at least I will never have to worry about going on a ventilator.

注意与下例进行区分：

[@Fever#Symptom*] has broke but have terrible [@**head and muscle ache**#Symptom*].，在该例子中，and 连接的不是症状而是身体部位，最终的症状是 ache，所以需要完整标注。

- 症状的一些**隐含表达**，也要标注完整。

示例
I asked about the vaccine but she said even though I have a [@**lung condition**#Symptom*] now
I feel [@**freezing from the inside out**#Symptom*] have a [@**100.5F temp**#Symptom*]
1st day of shot, felt [@**woozy**#Symptom*] with a slight [@**headache**#Symptom*]. 2nd day, mild [@**flu like symptoms**#Symptom*],
[@**Knocked me off my feet**#Symptom*].
I got [@**headaches**#Symptom*] after dose two but I also had [@**not slept well**#Symptom*] that night for other reasons.
i could [@**barely sleep all night**#Symptom*]
Had [@**fluey symptoms**#Symptom*] from jab but …
[@**flu-like symptoms**#Symptom*]
since then been v [@**sneezy**#Symptom*] and a bit [@**sniffly**#Symptom*].
my [@**head is killing**#Symptom*] me i hope this does not happen with the second dose of the covid vaccine
Day 1 after COVID vaccine and so far just a little more [@**tired than usual**#Symptom*], some mild body aches and a sore arm from the injection. Nothing compared to how awful I felt with COVID.
~21 hours after Dose 2 and I have a [@**100°F fever**#Symptom*], [@**cutaneous hyperesthesia**#Symptom*], [@**restless fatigue**#Symptom*].

- 隐含的 symptom 里的 feel like 也需要标注，否则会带来歧义。

示例
I woke up this morning [@**feeling like I have been hit by a truck**#Symptom*]
Even though I started the day out [@**feeling like I was hit by a Mack truck**#Symptom*].

- 其他容易漏标/错标的情况。

示例	说明
cough	网络用语，并非用来表示症状，故不用标注
With the **headache** of distribution on top of procurement, how will the world reach the herd immunity levels needed to defeat the virus?	此处 headache 并不表示症状，所以不应该标注
got the shot: [@**Pain**#Symptom*] [@**Swelling**#Symptom*] Throughout the rest of your body: [@**Fever**#Symptom*] [@**Chills**#Symptom*]	注意**不要**把 Throughout the rest of your body 也标注

	进去
How do they know if no test is available unless you [@**hv temp**#Symptom*], [@**cough**#Symptom*], or [@**loss of smell**#Symptom*]?	
but thank God I do not have the [@**chills**#Symptom*] and [@**cold sweats**#Symptom*] anymore!	
I got the second covid vaccine today I was totally prepared to be [@**exhausted**#Symptom*],	
if you add that to lockdown **fatigue**,	此处 fatigue 并不是来表示症状，故不应标注
Firstly, there is natural **fatigue** with Covid measures.	此处 fatigue 并不是来表示症状，故不应标注
Really did make me appreciate the face masks helping me avoid getting **sick** this last yea	此处 sick 泛指生病，故不应标注
QA! Can a COVID-19 vaccine make you **sick** with COVID-19? No, none of the vaccines in [@**the U.S.** #Location*] contain the love virus that causes COVID-19. The teaches your immune system how to recognize and fight the virus.	此处 sick 泛指生病，故不应标注
the second vaccine will probably make you **sick**.	此处 sick 泛指生病，故不应标注
at 3am woke up with [@**sweat on my forehead**#Symptom*]	
Well 1 day after the jab i felt [@**bloody awful**#Symptom*], [@**temperature**#Symptom*] all over the place n [@**headache**#Symptom*],	
I know some had more [@**pain**#Symptom*] with [@**chronic pain**#Symptom*] and worse [@**fatigue**#Symptom*] with [@**chronic fatigue**#Symptom*] Only asking for those with [@**chronic pain**#Symptom*] and/or [@**chronic fatigue**#Symptom*] who have had it.	
you will [@**not sleep**#Symptom*] tonight	
Much higher antibody titres but elevated rates of adverse effects too.	不属于症状，不应标注
light [@**headed**#Symptom*]/[@**dizzy a headache**#Symptom*].	
Now, left undesirable [@**lymph node is swollen**#Symptom*] and [@**fever**#Symptom*] is back.	
I never had [@**fever**#Symptom*] or any signs of [@**sickness**#Symptom*] after any vaccine I got so far,	
[@**Fever**#Symptom*] post vaccine? Do not **panic**, treat it!	panic 意思为慌张，不应标注为症状
Bit of a reaction to my Covid Vaccination. Woke up at 2am shivering my bits off with a [@**temperature of 39.4C**#Symptom*], [@**splitting headache**#Symptom*], [@**short of breath**#Symptom*] everything ached. Temp down to 38.6 but still got a [@**headache**#Symptom*], [@**wheezy a bit achey**#Symptom*]. I am doing nowt today, sofa, blanket hot choc!	

Arm had a [@**large, itchy, sore egg**#Symptom*] on the vaccination site which	
not only pandemic **fatigue**, but the inevitable selection of vaccine-escape variants if we keep cases high. and there lies	此处 fatigue 并不是症状，故不标注
unless you apply NPI, cases increase. but people are **tired**.	此处 tired 不属于症状，故不标注
I received “covid arm” and some lovely [@**fever blisters**#Symptom*]	
[@**Chronic COVID-19 Syndrome**#Symptom*] and **Chronic Fatigue Syndrome** following the first pandemic wave in [@**Germany**#Location*] – a first analysis of a prospective observational study	
WATCH: Dr. [@**Fauci**#Person*] discusses vaccine safety, COVID **fatigue** in appearance on News 12	此处 fatigue 不属于症状
Got mine two days ago and ran [@**fever**#Symptom*] and felt [@**awful all day**#Symptom*] yesterday.	
This time last year I went to a birthday party 2 days later I got [@**sick**#Symptom*]. At first I did not realize it was covid,	

5. 药物（DRU）

药物广义上是指用来预防、治疗及诊断疾病的物质，另外也包括临床诊断试剂。

在标注药物实体时，需要注意：

1. 药物的属性不可标注为“药物”
2. 大部分药物的统称均标注，比如比如营养素、抗菌药物、急救药物等，此类虽然是统称，但是有对应的治疗范畴。像“常用药”、“药物”等单独出现时，此类统称范围太广，不应标注。

- 容易错标/漏标的例子
  1. so hopefully that means it was not the [@**placebo**#Drug*].
  2. i used this same concoction others were mere [@**vitamin c**#Drug*], [@**zinc tabs**#Drug*] [@**azithromayicin**#Drug*] which is used to treat [@**cough**#Symptom*]

6. 疫苗相关（VAC）

包括疫苗名称、疫苗类型和疫苗品牌等。

附 1：新冠疫苗常见类型包括但不限于以下列表：

mRNA
protein subunit
subunit
vector
inactivated virus
peptide
synthetic
conjugate
plasmid

附 2：新冠疫苗常见品牌及其生产公司包括但不限于以下列表[2]：

疫苗品牌	别名	疫苗类型	生产商
Oxford–AstraZeneca	Vaxzevria	viral vector vaccine	British University of Oxford
	Covishield		British-Swedish company AstraZeneca
			Coalition for Epidemic Preparedness Innovations （CEPI）
Pfizer–BioNTech COVID-19 vaccine	Comirnaty	mRNA vaccine	German company BioNTech
			American company Pfizer
Moderna	COVID-19 Vaccine Moderna	mRNA vaccine	Moderna
			National Institute of Allergy and Infectious Diseases （NIAID）
			Biomedical Advanced Research and Development Authority （BARDA）
			Coalition for Epidemic Preparedness Innovations （CEPI）
Janssen	Johnson & Johnson	viral vector vaccine	Johnson & Johnson

[2] 该列表整理自 https://en.wikipedia.org/wiki/List_of_COVID-19_vaccine_authorizations

	COVID-19 Vaccine		
	COVID-19 Vaccine Janssen		Janssen Pharmaceutica
			Beth Israel Deaconess Medical Center （BIDMC）
Sinopharm-BBIBP	BBIBP-CorV	inactivated virus vaccine	China National Pharmaceutical Group Corporation （Sinopharm）
	Hayat-Vax		CNPGC
Sputnik V	Sputnik V COVID-19 vaccine	viral vector vaccine	Russian Gamaleya Research Institute of Epidemiology and Microbiology
CoronaVac		inactivated virus vaccine	Sinovac Biotech
Covaxin		inactivated virus vaccine	Bharat Biotech （BBIL）
			Indian Council of Medical Research （ICMR）
Sputnik Light		viral vector vaccine	Russian Gamaleya Research Institute of Epidemiology and Microbiology
Convidecia	Convidicea	viral vector vaccine	CanSino Biologics （CanSinoBIO）
			Beijing Institute of Biotechnology of the Academy of Military Medical Sciences （AMMS）
Sinopharm-WIBP	WIBP-CorV	inactivated virus vaccine	China National Pharmaceutical Group （Sinopharm）
			Wuhan Institute of Biological Products
EpiVacCorona		peptide vaccine	State Research Center of Virology and Biotechnology VECTOR
Zifivax		subunit vaccine	Anhui Zhifei Longcom Biopharmaceutical
Abdala		subunit vaccine	Center for Genetic Engineering and Biotechnology （CIGB）

Soberana 02		conjugate vaccine	Instituto Finlay de Vacunas
CoviVac		inactivated virus vaccine	Chumakov Centre at the Russian Academy of Sciences
QazCovid-in	QazVac	inactivated virus vaccine	Research Institute for Biological Safety Problems
Minhai		inactivated virus vaccine	Minhai Biotechnology Co.
			Shenzhen Kangtai Biological Products Co. Ltd.
COVIran Barakat		inactivated virus vaccine	Shifa Pharmed Industrial Co
Chinese Academy of Medical Sciences	Chinese Academy of Medical Sciences COVID-19 vaccine	inactivated virus vaccine	Chinese Academy of Medical Sciences
Medigen	MVC-COV1901	protein subunit vaccine	Medigen Vaccine Biologics and Dynavax Technologies
ZyCoV-D		DNA plasmid based COVID-19 vaccine	Cadila Healthcare
			Biotechnology Industry Research Assistance Council

注：疫苗的生产商应该标注为组织（ORG），疫苗带来的副作用标记为症状（SYM）。

- 当用生产商指代疫苗时，不应将生产商标注为组织（organization），而应该将其标注为疫苗相关。

示例
Alhamdulillah 4 of my family members had **the** [**@Pfizer vaccine#**Vaccine-related*]
Had the [**@Pfizer one#**Vaccine-related*]
Urging people to not get the [**@JJ vaccine#**Vaccine-related*] is crazy.

- 疫苗名+COVID-19 Vaccine 时应完整标注

示例
[**@Moderna COVID-19 Vaccine#**Vaccine-related*]
post 2nd dose of [**@Moderna Covid-19 vaccine#**Vaccine-related*]
the first [**@COVAX vaccine#**Vaccine-related*] shipment reaches [**@Ghana#**Location*],
[**@Pfizer covid-19 vaccine#**Vaccine-related*]

- **当出现疾病+疫苗时，也应标注为 vaccine-related[包括 COVID+vaccine 这**

样的表述]

Taken [@**Polio**, **Measles**, **Tetanus**, **Diphtheria vaccines**#Vaccine-related*] as a baby.

Received my 1st [@**Covid shot**#Vaccine-related*] yesterday.

I got the second dose of the [@**COVID19 vaccin**e#Vaccine-related*] yesterday.

I have had [@**BCG vaccine**#Vaccine-related*],[@**MMR vaccine**#Vaccine-related*] ,[@**Yellow fever vaccine**#Vaccine-related*],the [@**flu vaccine**#Vaccine-related*] ,and yesterday I had the [@**Pneumonia vaccine**#Vaccine-related*],

- 其他容易漏标、误标示例：

示例
地点+vaccine，则应将地点一并标注为 vaccine-related，即地点不再单独进行标注。 例如：in charge of distributing [@**West Virginia's vaccine**#Vaccine-related*] says it ma
Thanks and Was given [@**Oxford AstraZeneca vaccine**#Vaccine-related*].
do not use the [@**live virus**#Vaccine-related*] that causes COVID-19.
were noted in about 50% of participants in the [@**mRNA-1273**#Vaccine-related*] group after the second dose..
QA! Can a COVID-19 vaccine make you sick with COVID-19? No, none of the vaccines in [@**the U.S.**#Location*] contain the [@**love virus**#Vaccine-related*] that causes COVID-19. The teaches your immune system how to recognize and fight the virus.

7. 疾病（DIS）

导致病人处于非健康状态的原因或者医生对病人做出的诊断,并且是能够被治疗的，包括疾病或者综合征（disease or syndrome ）、中毒或受伤（injury or poisoning ）等。

类型	描述	示例
疾病或综合征	疾病或综合征是指疾病或综合征名称。	高血压（hypertension）、肺炎（pneumonia）、心脏病（heart disease）、败血症（sepsis）、畸形（deformity）等。
受伤或中毒	患者在受伤或中毒后,对人体造成某种伤害,导致患者处于非健康状态。	酒精中毒（alcoholism），食物中毒（sitotoxism）
器官或细胞损伤	器官、细胞等发生异常或损伤后,如果能够危及人的机体,此时虽然它们属于身体的一部 分,但是已成为一种致病因素,危害人体健康。	颅内出血（intracranial hemorrhage，ICH），产伤（birth trauma）

注：

1. 一般有些疾病名称很长,前面会有“XX 性”、“XX 状”、“XX 型”等,以及身体部位（一个或多个）的修饰,在保证疾病完整性和具体性的情况下,在标注时应该与这些前缀一起标注。
   示例：急性病毒感染（acute viral infection）、艾滋病毒急性感染 primary HIV infectio，急性黄疸型肝炎（Acute Icteric Hepatitis, AIH），乙型急性黄疸性肝炎（acute icteric hepatitis b），轮状病毒感染（Rotavirus infection），季节性婴幼儿腹泻（Seasonal infantile diarrhea），肝脏、肾脏、甲状腺疾病（Liver, kidney and thyroid diseases）
2. 大部分统称均标注,比如营养性疾病、代谢性疾病、化脓性和非化脓性综合征等,此类虽然是统称,但是有对应的疾病范畴,所以应该标注。特殊情况:像“常见病”、“ 多发病”、“疾病”等单独出现时,此类统称范围太广,不应标注。
   示例： [高血压]dis 是严重危害人类健康的常见病、多发病。
3. 当疾病有若干种分型 时,“疾病+分型”或“分型+疾病”整体标注,分型单独出现不标注。
   示例：II 型糖尿病（Type 2 diabetes）

- 注意一些用病毒名来指代疾病的例子，只有在确定病毒并非指代疾病时才不

用标注，否则一并按照指代疾病处理，标记为 Disease。

示例	说明
saying he had earned it over a lifetime of leadership on [@**HIV**#Disease*] research and [@**AIDS**#Disease*] relief	
Yep. [@**Zika**#Disease*]? Nothing burger; I	
[@**Ebola**#Disease*]? Boomslang snake and other reasons for [@**hemorrhagic fevers**#Disease*]	

- 一些容易错标、漏标的例子：

示例	说明
a regression w/o confounders like **obesity**, [@Azithromycin#Drug*], [@steroids#Drug*], [@comorbidities#Disease*],	obesity 表示肥胖，但是不一定是可以称之为疾病的肥胖
Felt like I had [@**flu**#Disease*]	只标注 flu 为疾病
it is more like [@**hayfever**#Disease*]	花粉热，一种疾病
Treated for [@**migraines**#Disease*] two days ago.	偏头痛，一种疾病
UNCLASSIFIED said mosquitoes not infected [@**w/yellow fever**#Disease*]	如果疾病和 w/之间没有空格，则一并标注，如果有空格则标注范围不包括 w/
[@TNXP#Organization*] connection to [@**Alzheimer**#Disease*]'s is pretty tenuous.	阿兹海默病
[@**TB**#Disease*] cert entering is for applicants staying 6 months	Tuberculosis，结核病的缩写
At least no more chance than the [@**common cold**#Disease*].	
[@**normal flu**#Disease*]	
Prior [@**SARS-CoV-2 infection**#Disease*]may have implications for vaccination approaches.	
I chose to have the [@**whooping cough**#Disease*] and [@**flu**#Disease*] vaccine in pregnancy	

8. @提及（Mention）标注策略

在英文推文中，@提及有时会作为推文正文的一部分，影响推文的语义。因此，可以进一步对参与句子表达的@提及进行标注，从而达到扩充 Person 及 Organization 两类实体样本量的目的。具体地，对满足以下条件的@提及进行标注：

1. 如果@提及出现在推文**首部**，则需判定在该@提及之后的**第一个单词**的首字母是否为小写，若为小写，则认为@提及属于句子一部分，故需要进行标注。反之无需标注；

例 1：@Pathik_Trader **in** some case rtpcr negative but covid positive if fever not reduce in 3 days then please test ct scan

【in 的首字母为小写，故应对@Pathik_Trader 进行标注】

例 2：@TheDavidL81 **Here's** the part where I explain Trump's strategy.

【Here 的首字母为大写，故**无需**对@TheDavidL81 进行标注】

2. 如果@提及出现在**正文**中，则需要对其标注。

例 3: The hypocrisy of @JoeBiden and @KamalaHarris is tragic.

【@JoeBiden 和@KamalaHarris 出现在句子中，因此要标注】

3. 如果@提及出现在推文**尾部**，则需查看该提及**之前的一个单词**，若为标签（#）或者标点符号，则无需进行标注，反之则认为其属于推文的一部分故需要对其进行标注。

例 4：Breathing difficulties with covid. #COVID @NHSuk

【@NHSuk 之前是标签，故**无需**标注】

例 5：it's #COVID19 coverage has been disgraceful. @BBCTalkback

【@BBCTalkback 之前是标点，故**无需**标注】

例 6：understand about COVID19. Follow @Cleavon_MD

【@Cleavon_MD 需要标注】

分类混淆处理

- 疾病（DIS）和症状（SYM）【更新时间：0827】
  疾病和临床表现的最大区别是:疾病是通过鉴别诊断的,疾病实质上就是身体受损;而症状实质上是身体受损后所表现出来的现象,比如说病人的不适感觉、身体出现的异常变化,但是这些往往是病人或者医生看到的表面现象。而作为医生则需要通过进一步的鉴别诊断来确认病人所患疾病,这也就说明疾病和临床表现存在着本质性的差异。
  在遇到“感染”相关的实体时,有以下几点需要注意一下:
  1. 若出现明确致病原因（病毒、细菌或身体部位等名称）与“感染”组合成词,则整体标注为“疾病”。如[HAV 感染]dis、[球菌感染]dis、[上呼吸道感染]dis
  2. 当单独出现“感染”一词时,若上下文明显表示是对某种疾病的指代,则标注为疾病,否则标注为“症状”
  3. 若“感染”一词前面的修饰词表明程度或频率时,则整体标注为“症状”

注：此部分根据标注者反馈不断更新。

标注质量评测

命名实体识别一般被建模成序列标注问题，因此常常使用微平均（Micro-averaging，Micro-F1）作为评价指标，计算公式如下：

$$P = \frac{识别正确的实体数目}{模型识别出的实体总数}$$

$$R = \frac{识别正确的实体数目}{参考结果中的实体总数}$$

$$Micro - F1 = \frac{2 \times P \times R}{P + R}$$