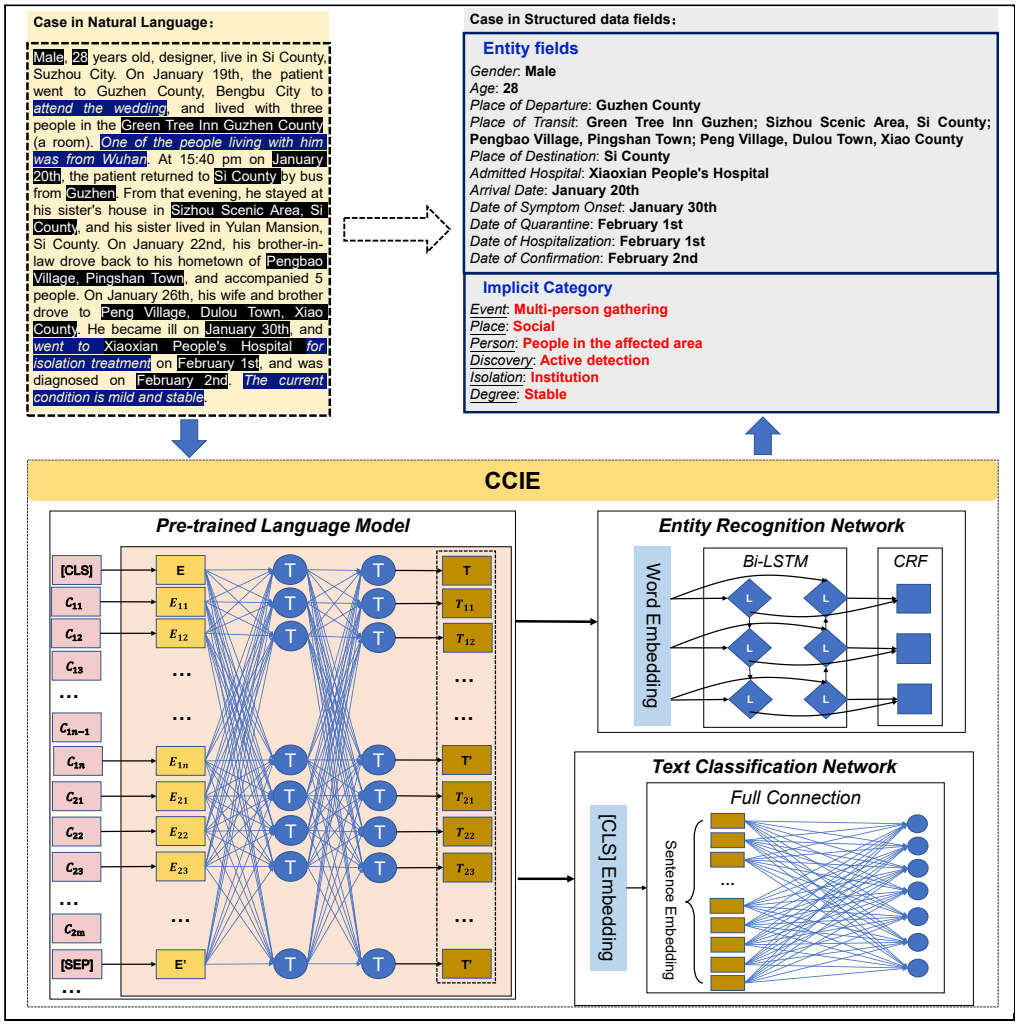
iScience

CellPress
OPEN ACCESS

Article

Epidemiologic information discovery from open-access COVID-19 case reports via pretrained language model



Zhizheng Wang, Xiao Fan Liu, Zhanwei Du, ..., Zoie S.Y. Wong, Xiao-Ke Xu, Yuanyuan Sun

zoiesywong@gmail.com (Z.S.Y.W.)
xuxiaoke@foxmail.com (X.-K.X.)
syuan@dlut.edu.cn (Y.S.)
lw660@cam.ac.uk (L.W.)

Highlights

We propose a method to obtain epidemiological information from COVID-19 case reports

The extracted information has 80% matching rate with the gold-standard manual coding

We provide an online platform that can analyze epidemiological statistics in real time

Wang et al., iScience 25, 105079
October 21, 2022
https://doi.org/10.1016/j.isci.2022.105079

iScience

CellPress
OPEN ACCESS

Article

Epidemiologic information discovery from open-access COVID-19 case reports via pretrained language model

Zhizheng Wang,[1,10] Xiao Fan Liu,[2,10] Zhanwei Du,[3,10] Lin Wang,[4,10,*] Ye Wu,[5] Petter Holme,[6] Michael Lachmann,[7] Hongfei Lin,[1] Zoie S.Y. Wong,[8,*] Xiao-Ke Xu,[9,*] and Yuanyuan Sun[1,11,*]

SUMMARY

Although open-access data are increasingly common and useful to epidemiological research, the curation of such datasets is resource-intensive and time-consuming. Despite the existence of a major source of COVID-19 data, the regularly disclosed case reports were often written in natural language with an unstructured format. Here, we propose a computational framework that can automatically extract epidemiological information from open-access COVID-19 case reports. We develop this framework by coupling a language model developed using deep neural networks with training samples compiled using an optimized data annotation strategy. When applied to the COVID-19 case reports collected from mainland China, our framework outperforms all other state-of-the-art deep learning models. The information extracted from our approach is highly consistent with that obtained from the gold-standard manual coding, with a matching rate of 80%. To disseminate our algorithm, we provide an open-access online platform that is able to estimate key epidemiological statistics in real time, with much less effort for data curation.

INTRODUCTION

The coronavirus disease 2019 (COVID-19) pandemic has been a global public health crisis (Malhotra et al., 2020; Le et al., 2020; Agbehadji et al., 2020; Chinazzi et al., 2020), with more than 300 million confirmed cases as of the end of 2021. Many countries and regions, such as China (Liu et al., 2021), Singapore (Singapore Ministry of Health, 2020), and Taiwan (Taiwan (2020), were able to publish COVID-19 case reports obtained from detailed epidemiological investigations in real time, with the aim of enhancing situation awareness (Bunker, 2020) and promoting individual behavior of self-protection (Zheng et al., 2021). These disclosed epidemiological survey results may contain demographic, travel-related, and diagnostic information for each confirmed case.

Analyses using open-access data have contributed key insights to help understand the epidemiological and pathological characteristics of COVID-19 (United States of America, 2021; Freunde von GISAID, 2021; Xu et al., 2020a; Du et al., 2020a; Ali et al., 2020; Xu et al., 2020b) to estimate the infection and disease burdens (O'Driscoll et al., 2021; Salje et al., 2020), characterize population behavioral changes (Zhang et al., 2020; Du et al., 2020b; Tian et al., 2020), and optimize control measures (Hale et al., 2021; Yang et al., 2021; World Health Organization, 2022). However, publicly disclosed case reports obtained from epidemiological investigations were often written in natural language without a standardized structure (e.g., different writers may use different words to express the same information). The data curation and standardization processes can be resource-intensive and time-consuming (Kraemer et al., 2021). For example, Liu et al. (Liu et al. (2021) recruited a team of 20 data curators to trace and manually annotate the demographic information, mobility history, and epidemiological timelines for each COVID-19 case that was publicly reported in mainland China as of March 4 2020. To reduce the burden on human resources and facilitate the real-time analyses of open-access case reports, the early research (Ghosh et al., 2017) inspired us to develop a deep learning framework using natural language processing (NLP) techniques to automatically identify key information from the raw case reports (Figure 1A).

Generally, different from the rule-based methods that use regularization formulation to match line lists from raw data, a deep learning framework that curates open case reports involves a combination of named

[1]College of Computer Science and Technology, Dalian University of Technology, Haishan Building No.2 Linggong Road, Dalian, Liaoning 116023, China

[2]Web Mining Laboratory, Department of Media and Communication, City University of Hong Kong, Hong Kong Special Administrative Region, China

[3]WHO Collaborating Centre for Infectious Disease Epidemiology and Control, School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong Special Administrative Region, China

[4]Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK

[5]Computational Communication Research Center and School of Journalism and Communication, Beijing Normal University, Beijing, China

[6]Tokyo Tech World Research Hub Initiative (WRHI), Institute of Innovative Research, Tokyo Institute of Technology, Tokyo, Japan

[7]Santa Fe Institute, Santa Fe, NM, USA

[8]Graduate School of Public Health, St. Luke's International University, Tokyo, Japan

[9]College of Information and Communication Engineering, Dalian Minzu University, Liaoning, China

[10]These authors contributed equally

[11]Lead contact

*Correspondence: zoiesywong@gmail.com (Z.S.Y.W.), xuxiaoke@foxmail.com (X.-K.X.), syuan@dlut.edu.cn (Y.S.), lw660@cam.ac.uk (L.W.)

https://doi.org/10.1016/j.isci.2022.105079

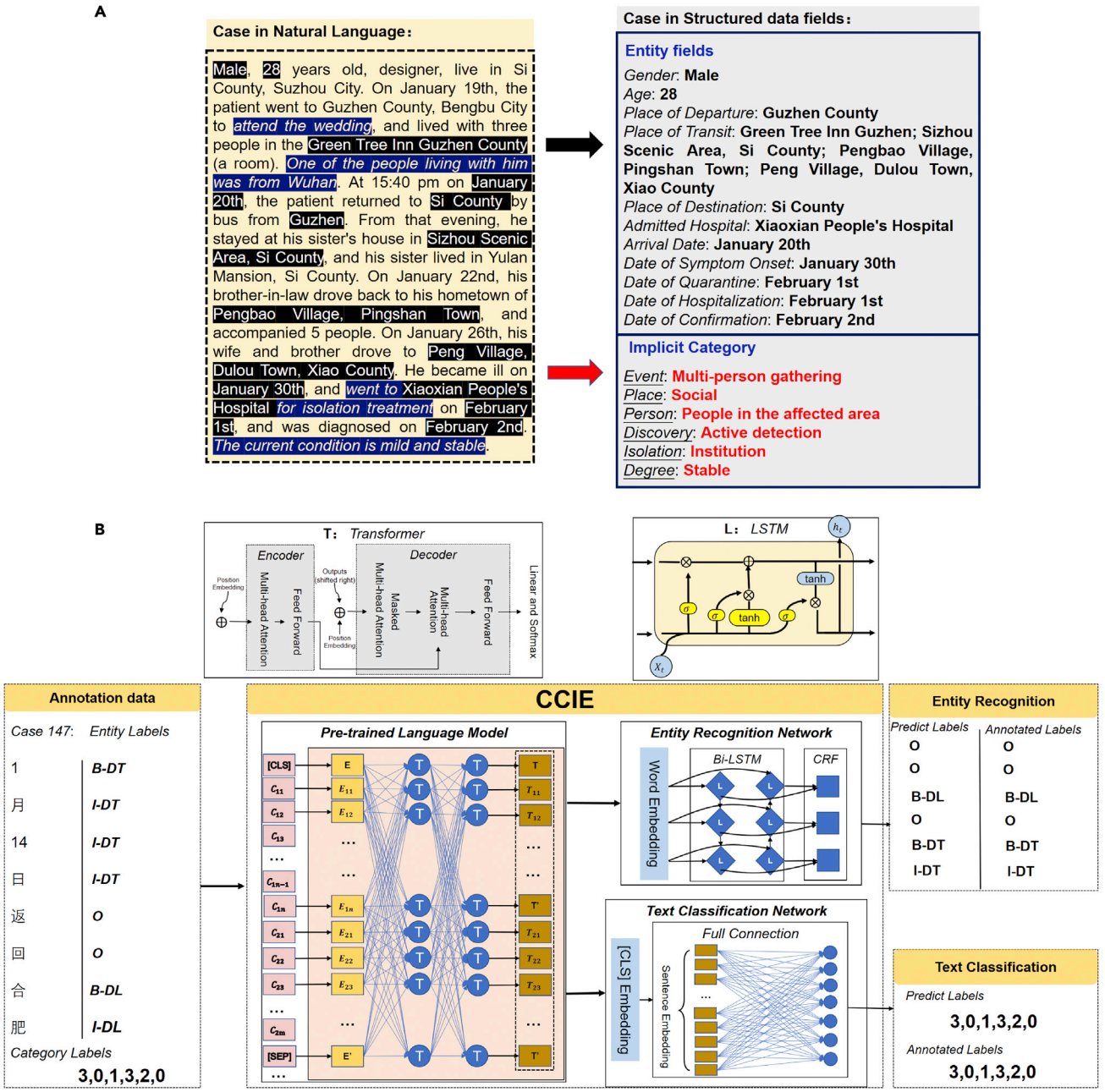


Figure 1. COVID-19 cases information extraction (CCIE) framework

(A) The CCIE framework can automatically translate data from open-access COVID-19 case reports into structured data fields.

(B) CCIE's workflow: The annotation data provided to the CCIE framework contain entity labels and category labels, with the letter "B-" indicating the start position of an entity and "I-" representing the middle or end position of an entity. CCIE comprises a pretrained language model, a named-entity-recognition network, and a text classification network. Specifically, the pretrained language model is built with *Transformer*, which uses each token of case reports as data input, with [CLS] indicating the start of a sentence and [SEP] denoting the separator between two adjacent sentences. The panel "T: Transformer" explains the internal structure of *Transformer*, with the symbol $\oplus$ indicating the concatenation operation. The named-entity-recognition network is built with the BiLSTM model and CRF predictor. The panel "L: LSTM" explains the internal structure of this network, with the symbol $\otimes$ indicating the elementwise multiplication, σ indicating the sigmoid function, *tanh* denoting the activation function, and X_t and h_t representing the input and output of *BiLSTM*, respectively. The text classification network consists of a fully connected neural network, with the [CLS] vector denoting sentence embedding. The assessment of model performance requires the evaluation of both the named-entity recognition and the text classification.

entity recognition, text classification, and knowledge inference tasks in NLP. For example, the identification of spatial locations and calendar dates requires named-entity recognition. Distinguishing the case detection method, such as the RT-PCR test or symptom onset, requires text classification, as the expressions often vary with the different natural language writing styles used in the reports. Where there were incomplete data fields or vague language expressions, knowledge inference and standardization were required (e.g., the correction of "Guzhen County" to "Suzhou City" according to geographical knowledge). The complexity in the real-world case report data prevents the direct application of advanced NLP tools, as exemplified by the poor performance of applying the seminal pretrained language model (Devlin et al., 2019) directly to the human-coded data (Figure S1).

Therefore, we were required to adjust the existing NLP models for our data curation task via the preparation of appropriately high-quality annotated data. We first propose a machine-learnable annotation strategy to refine the codebook in ref. (Liu et al., 2021), in which we target 17 data fields and group them into named-entity-recognition tasks and text classification tasks. After that, we propose the COVID-19 cases information extraction (CCIE) framework, which uses three deep neural networks to perform the named-entity recognition and text classification (Figure 1B). The pretrained language model with the whole-word-mark (WWM) mechanism (Cui et al., 2021) encodes each case report into vector representations; a bidirectional long short-term memory (Bi-LSTM) (Kadari et al., 2017) performs the named-entity recognition, and a fully connected network performs the text classification. Finally, we evaluate CCIE in relation to three aspects. First, we apply our annotation strategy to different deep neural networks to observe the adaptability of the annotated data. Second, we compare the proposed CCIE framework with state-of-the-art models with regard to the tasks of named-entity recognition and text classification. Last, we investigate the effectiveness of the CCIE framework through cross-validation using manually extracted values. In practice, we also develop an online system based on CCIE, which is publicly available to all researchers worldwide.

RESULTS

Performance evaluation

Annotation strategy

The annotations of case reports were used as labels for different deep neural networks. To guarantee the consistency and accuracy of the manual annotation, we randomly examined and modified a subset of 100 case reports after they had been annotated by different graduate students. Then, three public-health experts participated in the revisions. After these, we discussed the annotations of the case reports to reach a consensus on the modifications. Next, we continued to examine and manually annotate the remaining case reports. The agreement rate for our revisions reaches 90%, which suggests that the inter-annotator agreement rate is acceptable. Any inconsistent revisions were submitted to the experts for final revisions. Based on our annotation data, all the deep neural networks demonstrated high adaptability across different tasks (i.e., named-entity recognition and text classification). For example, in the named-entity-recognition tasks (Figure 2A), all models achieved F1 values of higher than 70% for most entities. For fields with a fixed language format, such as "dates," and obvious trigger words, such as *admitted hospital*, the F1 values (a global evaluation for precision and recall, which were calculated using Equation (9)) for all models exceed 90%. Notably, for fields with a long text length and high ambiguity, such as "place of transit," most deep neural network models obtained F1 values of over 50%. In the text classification tasks (Figure 2B), most models achieved F1 values of higher than 80% for the data fields with limited labels. Even for the category with more possible labels, such as "event," the evaluated models obtained an F1 value that exceeded 75%. All these results passed the t-test, with a confidence interval of 0.95, under the assumption that all values follow normal distribution.

Text classification tasks

To further analyze the performance of the CCIE, we first compared it with seven benchmark text classification algorithms (i.e., Transformer (Vaswani et al., 2017), DPCNN (Johnson and Zhang, 2017), FastText (Joulin et al., 2017), TextCNN (Kim, 2014), TextRNN (Liu et al., 2016), TextRCNN (Lai et al., 2015), and LSTM (Zhou et al., 2016)) for the classification of six categories (Table S1A). The F1 values obtained by the CCIE for all six categories were above 82%, with the highest value reaching 93.2%. Especially in the *event* category, with maximum category labels, the CCIE increased by 2.6% compared to TextCNN and by 10.7% compared to Transformer. This result shows that the pretrained language models obtained word embeddings with richer semantic expression for mining deep features in the text, such as syntactic dependence and semantic role.

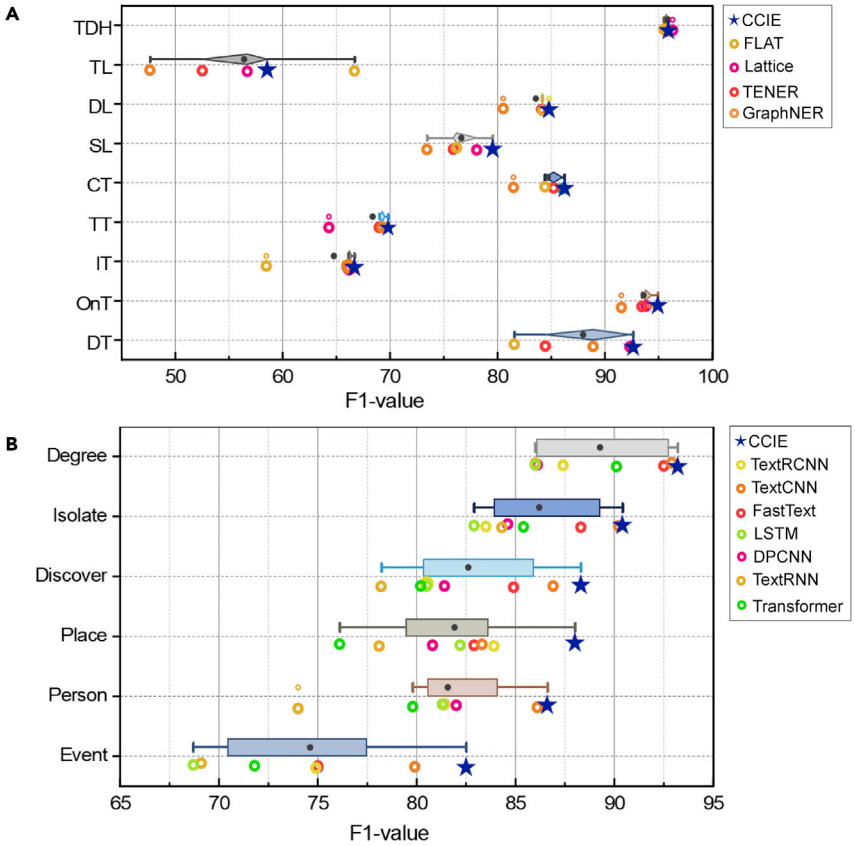


Figure 2. Distributions of F1 values for the named-entity-recognition and text classification models, obtained using our proposed annotation strategy

(A) The distribution of the F1 value for each named entity, which aggregates the results of five different named-entity-recognition methods, namely Lattice (Zhang and Yang, 2018), TENER (Yan et al., 2019), GraphNER (Sui et al., 2019), FLAT (Li et al., 2020), and our CCIE (denoted as "★"). The colored distributions correspond to different named entities.

(B) The distribution of the F1 value for each text category, which aggregates the results of eight text classification methods, namely Transformer (Vaswani et al., 2017), DPCNN (Johnson and Zhang, 2017), FastText (Joulin et al., 2017), TextCNN (Kim, 2014), TextRNN (Liu et al., 2016), TextRCNN (Lai et al., 2015), LSTM (Zhou et al., 2016), and our CCIE. For each named entity or category, the scattered dots indicate the F1 values obtained from different methods, which are used to fit the distribution as indicated by the box plot. [See also Figure S1 and Tables S1 and S6].

Named-entity recognition tasks

Then, we also compared the CCIE with four classic deep neural network models (i.e., Lattice (Zhang and Yang, 2018), Tener (Yan et al., 2019), GraphNER (Sui et al., 2019), and FLAT (Li et al., 2020)) with regard to the recognition of nine entities (Table S1B). The CCIE demonstrated better performance for most entity recognitions (7/9 cases), including all "dates" and two "places." This achievement is attributed to the fact that the pretrained model used for the CCIE adopts the WWM mechanism to capture the regular date format "xx (month) xx day, xx year", or "xx (month) xx day" and helps determine the entity boundaries. The same reason applies to the recognition of the *departure place* and the *destination place*, as the description granularities of these fields are recorded as "xx City" and "xx County," respectively. In addition, we compared the CCIE with other pretrained language model-based solutions (i.e., "TENER + BERT" and "TENER + (BERT with WWM)"). The results show that the CCIE outperforms "TENER + BERT" and obtains results comparable to those of "TENER + (BERT with WWM)" (Table S1C), which, in turn, indicates that the WWM mechanism is key for identifying entity boundaries.

Sample size threshold of annotated data

Given that data annotation requires considerable labor, determining the minimum label set size for the models to obtain reasonable performance is important. Therefore, we conducted the named-entity-recognition task

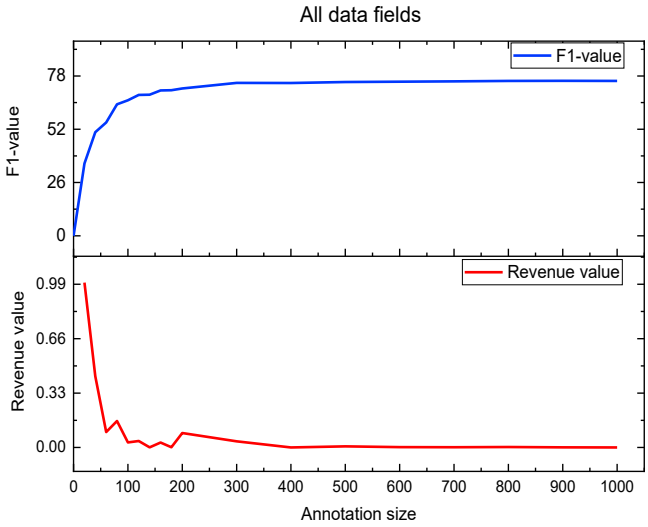


Figure 3. Performance gain across all data fields by increasing the annotation size

The CCIE framework was used for the named-entity recognition and text classification. The blue curve in the upper panel indicates the increase in the overall F1 value as the size of the annotation increases. The red curve in the bottom panel indicates the reduction in the revenue value as the size of the annotation increases. The revenue value is calculated as *(after-orginal)/original*; *after* indicates the F1 value of CCIE after the size of the annotation data is increased, while *original* indicates the F1 value of the CCIE before the addition of annotation data. When the size of the annotation data is smaller than 200, the revenue value is calculated for every additional 20 annotation data; further, when it is larger than 200, the revenue value is calculated for every additional 100 annotation data. [See also Figure S2].

using the CCIE under different annotated data volumes. With an increase in the annotated data size, the overall performance for named-entity recognition showed an upward trend (Figure 3). However, when the annotated data size reached 400, the upward trend was no longer evident, and the revenue curve appeared to be stable between 0.1% and 0.2%. Moreover, from the perspective of the recognition accuracy for each entity (Figure S2), the result and revenue stopped fluctuating significantly when the annotated data size reached 600, and the average revenue remained between 0% and 0.04%. Note that the recognition of *admitted hospital* reached an optimum value (90%) when the annotated data size was merely 100, meaning that the more evident the trigger word is, the lesser data are required to be annotated.

Performance variance due to language styles

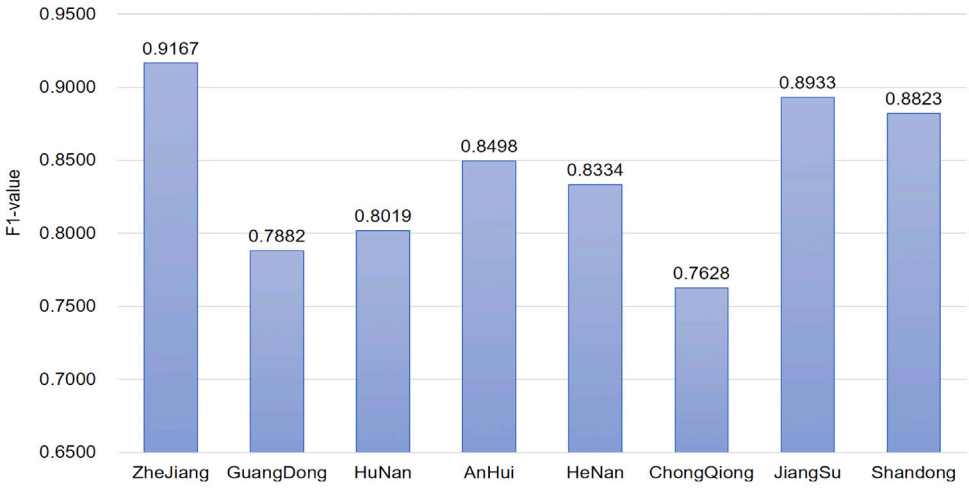
The confirmed COVID-19 cases in our dataset were reported from 27 provincial health departments and 264 municipal health departments, which lead to large differences in their language styles. Moreover, our CCIE is a feature learning model and, therefore, sensitive to language styles. We selected eight provinces that reported the highest number of cases (i.e., Zhejiang, Jiangsu, Shandong, Guangdong, Chongqing, Hunan, Anhui, and Henan) and compared the CCIE performance for the reports released by each province (Figure 4A). The CCIE framework performed well for the reports released by the health departments of Zhejiang (91.67%), Jiangsu (89.33%), and Shandong (88.23%) but not for those released by the health departments of Guangdong (78.82%) and Chongqing (76.28%).

After parsing the report examples released by different provinces (Figure 4B), we found that case reports can be most easily processed by the CCIE when (1) the reported entities have a concise text description, (2) the correspondence between the trigger words and entities is clear and unique, and (3) the distance between an entity and its trigger words is relatively short. Therefore, we propose a template for future epidemiology surveys (Table S2) and design the corresponding questions that should be asked in epidemiology surveys (Table S3), which covers travel history and social (contact) behaviors.

Cross-validation with manually extracted information

We compared the 17 machine-extracted fields with the manually extracted ones from the dataset of Liu et al. (Liu et al. (2021) for the first 10,000 case disclosure reports (i.e., from January 2 to March 4, 2020). We used a simple fuzzy matching logic (Figure 5A) to deal with the style differences between the machine- and

A



B

Province	COVID-19 Case Report
ZheJiang	Male, 41 years old, living in [Luqiao District (***OTL***)]. He developed a fever on [January 15th (***OnT***)], and the body temperature is 38.1° C, chills, coughs, accompanies fatigued muscle aches. He came to [Taizhou (***DL***)] from [Wuhan (***SL***)] on [January 21st (***DT***)] for medical treatment, and now he is in [Taizhou City Health medical center (***TDH***)] for treatment.
GuangDong	Male, 40 years old, resident in [Wuhan City, Hubei Province (***OTL***)]. On the [evening of January 21 (***DT***)], take a train from [Wuhan (***SL***)] to [Guangzhou Station (***TL***)], then transferred to Guangzhou-Zhuhai Railway Station to [Zhuhai Station (***D***)] on the [22nd (***DT***)], and then check into the hotel at [Hengqin (***TL***)]. He accepted medical observation at the hotel from the [24th (***IT***)]. On the [30th (***CT***)], the nucleic acid test was positive, and he was transferred to the [fifth hospital of CUHK (***TDH***)] for treatment.
ChongQing	Male, 45 years old, liveing in [Jianlou Village, Tiefeng Township, Wanzhou District (***OTL***)] with his mother Yang Mou (the 45th confirmed case in [Wanzhou District (***OTL***)]). He attended to the wedding banquet at Liu's house on [Tiefeng New Street (***TL***)]. At [7 o'clock on the 21st (***OTT***)], he take the shuttle bus from Tiefeng to [Little Swan (***TL***)] to buy new year's goods, then take a taxi to [a pharmacy in Shanshui International (***TL***)] to buy medicine, and then ate at [a restaurant in Shanshui International (***TL***)], then take the shuttle back to [Tiefeng Township (***DL***)]] at [14:00 (***DT***)]. Cough occurred on [February 1 (***OnT***)] and was controlled as a close contact. The improvement was not obvious after the medicine was delivered by the [Tiefeng Township Health Hospital (***TDH***)]. At [12 o'clock on February 5th (***IT***)], the [First People's Hospital of Wanzhou District (***TDH***)] picked him up. The diagnosis was confirmed after examination on [February 6 (***CT***)].
JiangSu	Female, 50 years old, currently living in [Hehu Xingcheng Community, Huai'an District (***OTL***)], with no history of [Hubei (***OTL***)] residence and travel. Her son was in contact with a confirmed patient in our city. On [February 3rd (***OnT***)], the patient developed symptoms such as sore throat. She rested after taking the medicine on his own and the symptoms were relieved; On [February 5th (***OnT***)], symptoms such as fever and joint soreness occurred; On [February 6th (***TT***)], she went to [Huai'an Hospital in Huaian City (***TDH***)] for treatment, and then transferred to the [Fourth People's Hospital of Huai'an City (***TDH***)]; [February 13rd (***CT***)] was diagnosed with new coronary pneumonia.
ShanDong	Male, 57 years old, native of [Banquan Town, Junan County (***OTL***)], works in [Wuhan (***OTL***)]. On [January 21 (***DT***)], he returned to [Junan (***L***)] from [Wuhan (***SL***)] by car. On [January 29 (***OnT***)], he developed symptoms such as diarrhea, which were transferred to and treated by [designated medical institutions (***TDH***)]. After [laboratory (***OTl***)] testing, the new coronavirus nucleic acid was positive, and the condition is stable.

Figure 4. Performance of CCIE in terms of identifying data fields from the COVID-19 case reports disclosed by each province in China

(A) Eight provinces, Zhejiang, Guangdong, Hunan, Anhui, Henan, Chongqing, Jiangsu, and Shandong, were selected to assess the effectiveness of the CCIE with regard to the handling of case reports with different natural-language writing styles.

(B) To illustrate the data fields identified using our CCIE framework, the text boxes provide five examples of case reports from Zhejiang, Guangdong, Chongqing, Jiangsu, and Shandong, with the identified data fields highlighted in blue, the trigger words highlighted in red, and the field labels highlighted in bold black. [See also Tables S2 and S3].

human-extracted fields—if the machine-extracted text was present in the manually coded fields, we considered the machine to have provided meaningful information. The comparison results (Figure 5B) showed high consistency between machine extraction and manual coding. The agreement rates for *ages* and *genders* were 97.2% and 97.96%, respectively; those for *the places of departure, transit,* and *destination* were 74.95%, 86.84%, and 66.62%, respectively; and those for the *dates of arrival, quarantine, symptom onset, hospitalization,* and *confirmation* were 87.67%, 75.14%, 86.21%, 69.24%, and 65.89%, respectively.

The matching rate for the *admitted hospital* field was relatively low. We found that the inconsistency between machine extraction and human coding stems from the following facts: (i) The machine extracts the abbreviations of hospital names, while human coding converts them into full names (~85% of the

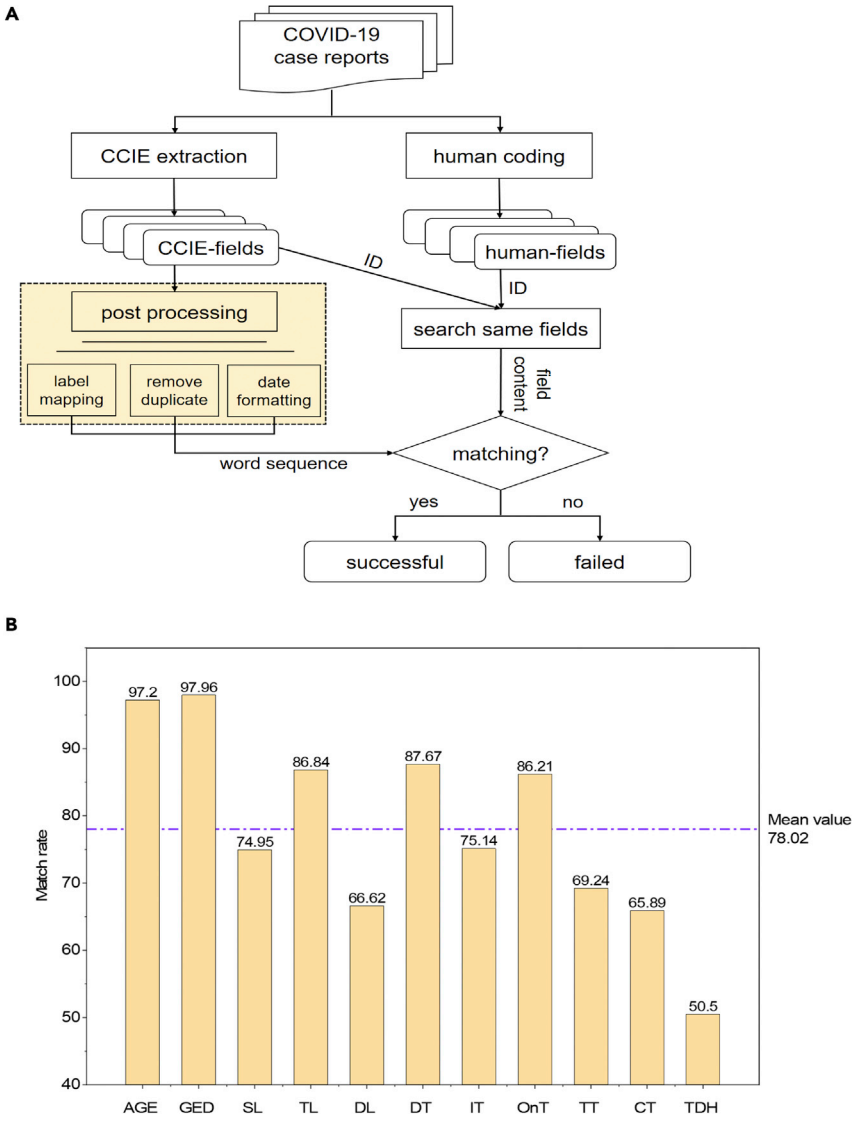


Figure 5. Cross-validation with manually extracted named entities

(A) The fuzzy matching method (highlighted in yellow) is used to compute the matching rate, where the term "label mapping" indicates the projection of the named entities obtained from our CCIE into the readable fields, "remove duplicate" indicates the removal of duplicate values for the same data field, and "date formatting" indicates the calibration of the date fields identified from our CCIE into the regular format of "xx (month) xx-day, xx year."

(B) The accuracy of the eleven data fields identified using our CCIE algorithm as compared to the gold-standard results obtained by manual human coding; the mean value is obtained by averaging the accuracy of all data fields.

cases); (ii) the vague term "designated medical institution" used by the local authorities instead of the exact hospital names was not considered in human coding but recognized by the machine (~10% of the cases); and (iii) the machine failed to recognize the correct *admitted hospital* field mentioned in the report or did not recognize this element at all (~5% of the cases).

The machine-extracted information could also be used to determine the epidemiological characteristics of COVID-19 with close-to-human-coding precision. For example, we used *the date of the symptom onset* field to compute the real-time reproduction (R_t) number. From January 8 to February 26 2020, the distribution of R_t values calculated by human coding and machine extraction remains consistent. We also calculated the R-square value and root-mean-square error (RMSE) for these two groups of numerical distributions (Figure 6). The results demonstrated high consistency between the two distributions, which

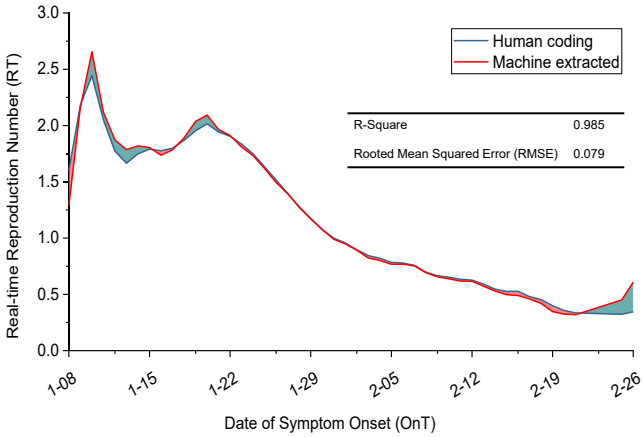


Figure 6. Confluence between the real-time reproduction number (R_t) estimated using data fields of symptom onset date identified from our CCIE and that estimated using gold-standard data produced through manual human coding

The analysis used COVID-19 cases with symptom onset occurring between January 8 and February 26 2020. The R_t was estimated using a ready-to-use tool (Cori et al., 2013), which was implemented in popular software including Microsoft Excel. To quantify the accuracy of the CCIE with regard to estimating R_t, the inset panel shows the R-square and the root-mean-square error (RMSE) for the time series of R_t, which was estimated using two data-extraction methods.

showed that the field extracted by the machine yields a result that is comparable to that obtained through human encoding in the calculation of the R_t index.

Details of the online system

We provided an online system (*http://covid19.caseassistant.top*) to help extract structured data fields from open-access COVID-19 case reports. The system can automatically extract the activity trajectory (e.g., *places of departure*, *transit*, and *destination*), infection cycle (such as *dates of arrival*, *symptom onset*, *quarantine*, *hospitalization*, and *confirmation*), and the *admitted hospital* of infected patients. We organized the location fields extracted from an infection case into a timeline based on temporal logic to allow researchers to more intuitively grasp the activity trajectory of infected patients. We also added a geographical analysis module of infection cases to the system—which can count the high incidence areas of COVID-19 according to the location of the infected person—to analyze the geographical distribution of disease transmission in a targeted manner. The system exhibits high scalability and can satisfy the deployment of both GPU and CPU environments simultaneously. The average processing speed for the GPU is five seconds per case, while that for the CPU is approximately ten seconds per case.

DISCUSSION

The epidemiological analysis of community transmission is vital for formulating public health interventions against COVID-19 (Byambasuren et al., 2020; Whaiduzzaman et al., 2020). This is critical for clarifying the host selection and physiological mechanism of COVID-19, as it allows us to obtain essential content, such as the gathering behavior and activity trajectory for the massive infection cases. To facilitate the automatic extraction of epidemiological information from open-access COVID-19 case reports, we first proposed a refined annotation strategy using the available human coding and then developed an information extraction framework that incorporates multiple deep neural networks to perform the named-entity recognition and text classification tasks. The accuracy of our CCIE framework is very high (>80%), which outperforms several state-of-the-art models such as Transformer (Vaswani et al., 2017), LSTM (Zhou et al., 2016), Lattice (Zhang and Yang, 2018), and TENER (Yan et al., 2019). In particular, our method reduces, on average, around 80% of the labor (about 20 annotators), who work on the manual coding of raw case reports written in natural language; in addition, the machine-extracted data fields are able to correct some fields that were incorrectly coded by humans, such as the inconsistency in the word segments extracted for *admitted hospital*.

To ease the implementation of our framework, we provided an online system that can be accessed through this website: http://covid19.caseassistant.top. This system allows users to extract all 17 data

fields analyzed in our study from their respective case reports. This serves as a preliminary step in the automatic information extraction of epidemiology survey reports and is expected to benefit the wider research community.

Our system automatically extracts key epidemiological information, including demographics, travel history, contact scenarios, and epidemiology timeline information, from the open-access case reports and has great potential to accelerate COVID-19 research. Although we only focus on the case reports written in Chinese here, our CCIE framework can be easily adapted for other languages. This is because our annotation strategy can be used for case reports written in different language styles, and we can easily change the pretrained language model used for Chinese to the most suitable models for a different language.

However, caution is needed when attempting to apply our framework to other situations. This may require a clear understanding of the background information. For example, raw case reports might contain a sentence like "the patient showed symptom on January 25 and was sent to hospital on 26." Our algorithm will not be able to extract "January 26" as the "hospitalization date", because of the lack of indication that the number 26 actually denotes the calendar date. Same problem may exist when extracting the "hospitalization date" and "confirmation date". Although our framework can extract the "admitted hospital" from case reports, it may identify an improper hospital if some patient transferred among multiple hospitals before the final admission. Nonetheless, these problems can be resolved with a more comprehensive annotation strategy, such as with additional definitions of the attributes and relations to describe the relationship between words (Brat nlplab, 2020).

Therefore, we call for standardization of the case reporting format and propose additional questions that should be asked in epidemiology surveys (Table S3), covering travel history and social (contact) behaviors. In particular, we distinguish the respondents based on whether they belong to returning home from abroad, which dissolves the information diversion in the case release. The content involved in the questionnaire refers to the publicly released report without any personal privacy, and our design makes it closer to the format that the NLP algorithm can directly handle. Compared to the traditional epidemiological questionnaire (Beijing Preventive Medicine Association, 2020), our designed questionnaire focuses on the trajectory of the infected person and the exact dates, which compensate for the information absence of infection cases. In addition, the questions and options of this questionnaire are fault-tolerant to a certain extent, which can accommodate the respondents' understanding of specific questions. Thereby, it effectively reduces the difficulty of information processing after the data collection.

Rapid COVID-19 linelist data curation and sharing have been emphasized by public health organizations and research institutions from the start of the COVID-19 pandemic (Moorthy et al., 2020). Although there are exemplar communities (GlobalHealth, 2022) hosting data repositories, the lack of structure hinders data processing at a large scale (Gardner et al., 2021). The raw COVID-19 linelist data from official case reports is unstructured data with application limitations. Analysis of such unstructured data is very complicated and slow. Although deep learning models have a great potential for learning the complex rules underlying the case reports, there is no study trying to extract structural fields from raw COVID-19 case reports. Our work contributes to automated data extraction and can be easily extended to data structured processing of publicly available unstructured data, which is attributed to the flexibility of neural network models. Making these data easy to use can not only mobilize interested researchers but also saves their effort in going through lengthy ethical review process before obtaining data for their studies. What's more, our work will continuously serve for curating the new COVID-19 case reports of mainland China.

Ethical approval for this study was provided by the Ethics Committee of Dalian University of Technology (Approval code: DUTIEE220615_01). During the data collection process, we followed the usage guidelines of the data publishing platform and utilized the collected data only for scientific research; further, the data were obtained with the consent of all participants. In the data processing stage, we sincerely considered the ethical decision making of regulatory bodies, strictly abided by ethical regulations to protect all private information in the data, and established an external advisory committee to supervise the data processing activity. However, we have to admit that, when disseminating

the data and technology, there will be challenges of information leakage and technology abuse. Therefore, we will strive to improve the security of the information, control the motivations for the continued use of technology, and improve the timeliness of decision feedback. We will design a more complete registration mechanism to take into account the motivation of the platform users. When publishing the data, the anonymized information in the data to be disclosed will be strictly screened to protect data privacy.

Limitations of the study

We implement automatic extraction of information from COVID-19 case reports using natural language processing techniques. Such a technical solution currently cannot fully identify specialized information that requires background knowledge and reasoning. Although we have adopted the optimal system deployment scheme and privacy protection mechanism, the operating efficiency of online systems under different network environments and hardware configurations is still unclear.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Data preprocessing
  - Annotation strategy
  - Structure of CCIE
  - Model training
  - Parameter setting
- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105079.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China [61773091 and 62173065 to X.-K.X., 72025405, 82041020 and 91846301 to X.L.]; Liaoning Revitalization Talents Program [XLYC1807106 to X.-K.X.]; Grand Challenges ICODA pilot initiative, delivered by Health Data Research UK and funded by the Bill & Melinda Gates Foundation and the Minderoo Foundation [to X.F.L.]; Japan Society for the Promotion of Science KAKENHI [Grant No. 18H03336 to Z.S.Y.W.]; and the Fundamental Research Funds for the Central Universities [No. DUT22ZD205 to Y.Y.S.].

AUTHOR CONTRIBUTIONS

Methodology: X.-K.X., Y.Y.Y.S.; Investigation: Z.W., X.F.L., Z.D., L.W.; Visualization: Z.W., Z.D.; Funding acquisition: X.K.X., L.W., X.F.L., Z.S.Y.W., Y.S.; Supervision: Y.W., P.H., M.L., H.L., Z.S.Y.W.; Writing – original draft: Z.W., X.F.L., Z.D., L.W.; Writing – review & editing: Z.W., X.F.L., Z.D., Z.S.Y.W., X.-K.X., Y.Y.S.

DECLARATION OF INTEREST

All authors declare no competing interests.

Received: March 12, 2022
Revised: July 4, 2022
Accepted: August 31, 2022
Published: October 21, 2022

REFERENCES

Agbehadji, I.E., Awuzie, B.O., Ngowi, A.B., and Millham, R.C. (2020). Review of big data analytics, artificial intelligence and nature-inspired computing models towards accurate detection of COVID-19 pandemic cases and contact tracing. Int. J. Environ. Res. Public Health *17*, 5330. https://doi.org/10.3390/ijerph17155330.

Ali, S.T., Wang, L., Lau, E.H.Y., Xu, X.K., Du, Z., Wu, Y., Leung, G.M., and Cowling, B.J. (2020). Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions. Science *369*, 1106–1109. https://doi.org/10.1126/science.abc9004.

Beijing Preventive Medicine Association (2020). Guideline for epidemiological investigation of coronavirus disease 2019 (T/BPMA 0003-2020). Zhonghua Liuxingbingxue Zazhi *41*, 1184–1191. https://doi.org/10.3760/cma.j.cn112338-20200421-00607.

Brat nlplab (2020). Brat Rapid Annotation Tool Manual. http://brat.nlplab.org/manual.html.

Bunker, D. (2020). Who do you trust? The digital destruction of shared situational awareness and the COVID-19 infodemic. Int. J. Inf. Manage. *55*, 102201. https://doi.org/10.1016/j.ijinfomgt.2020.102201.

Byambasuren, O., Cardona, M., Bell, K., Clark, J., McLaws, M.-L., and Glasziou, P. (2020). Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: systematic review and meta-analysis. Official Journal of the Association of Medical Microbiology and Infectious Disease Canada *5*, 223–234. https://doi.org/10.3138/jammi-2020-0030.

Chinazzi, M., Davis, J.T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., Pastore Y Piontti, A., Mu, K., Rossi, L., Sun, K., et al. (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science *368*, 395–400. https://doi.org/10.1126/science.aba9757.

Cori, A., Ferguson, N.M., Fraser, C., and Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. Am. J. Epidemiol. *178*, 1505–1512. https://doi.org/10.1093/aje/kwt133.

Cui, Y., Che, W., Liu, T., Qin, B., and Yang, Z. (2021). Pre-training with whole word masking for Chinese bert. IEEE/ACM Trans. Audio Speech Lang. Process. *29*, 3504–3514. https://doi.org/10.1109/TASLP.2021.3124365.

Devlin, J., Chang, M.W., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT *2019*, 4171–4186. https://doi.org/10.48550/arXiv.1810.04805.

Du, Z., Wang, L., Cauchemez, S., Xu, X., Wang, X., Cowling, B.J., and Meyers, L.A. (2020a). Risk for transportation of coronavirus disease from wuhan to other cities in China. Emerg. Infect. Dis. *26*, 1049–1052. https://doi.org/10.3201/eid2605.200146.

Du, Z., Xu, X., Wang, L., Fox, S.J., Cowling, B.J., Galvani, A.P., and Meyers, L.A. (2020b). Effects of proactive social distancing on COVID-19 outbreaks in 58 cities, China. Emerg. Infect. Dis. *26*, 2267–2269. https://doi.org/10.3201/eid2609.201932.

Freunde von GISAID, e.V. (2021). GISAID. https://www.gisaid.org/.

Gardner, L., Ratcliff, J., Dong, E., and Katz, A. (2021). A need for open public data standards and sharing in light of COVID-19. Lancet Infect. Dis. *21*, e80. https://doi.org/10.1016/S1473-3099(20)30635-6.

Ghosh, S., Chakraborty, P., Lewis, B.L., Majumder, M.S., Cohn, E., Brownstein, J.S., Marathe, M.V., and Ramakrishnan, N. (2017). GELL: automatic extraction of epidemiological line lists from open sources. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining *2017*, 1477–1485. https://doi.org/10.1145/3097983. 3098073.

GlobalHealth. (2022). A Data Science Initiative. https://global.health/.

Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., Webster, S., Cameron-Blake, E., Hallas, L., Majumdar, S., and Tatlow, H. (2021). A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). Nat. Hum. Behav. *5*, 529–538. https://doi.org/10.1038/s41562-021-01079-8.

Hu, W. (2020). Chinese Text Classification. https://github.com/649453932/Chinese-Text-Classification-Pytorch.

Johnson, R., and Zhang, T. (2017). Deep pyramid convolutional neural networks for text categorization. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics *2017*, 562–570. https://doi.org/10.18653/v1/P17-1052.

Joulin, A., Grave, É., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics *2017*, 427–431. https://doi.org/10.48550/arXiv.1607.01759.

Kadari, R., Zhang, Y., Zhang, W., and Liu, T. (2017). CCG supertagging via Bidirectional LSTM-CRF neural architecture. Neurocomputing *283*, 31–37. https://doi.org/10.1016/j.neucom.2017.12.050.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)), pp. 1746–1751. https://doi.org/10.48550/arXiv.1408.5882.

Kraemer, M.U.G., Scarpino, S.V., Marivate, V., Gutierrez, B., Xu, B., Lee, G., Hawkins, J.B., Rivers, C., Pigott, D.M., Katz, R., and Brownstein, J.S. (2021). Data curation during a pandemic and lessons learned from COVID-19. Nat. Comput. Sci. *1*, 9–10. https://doi.org/10.1038/s43588-020-00015-6.

Lai, S., Xu, L., Liu, K., and Zhao, J. (2015). Recurrent convolutional neural networks for text classification. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence 2015, 2267–2273. https://doi.org/10.1609/aaai.v29i1.9513.

Thanh Le, T., Andreadakis, Z., Kumar, A., Gómez Román, R., Tollefsen, S., Saville, M., and Mayhew, S. (2020). The COVID-19 vaccine development landscape. Nat. Rev. Drug Discov. *19*, 305–306. https://doi.org/10.1038/d41573-020-00073-5.

Li, X., Yan, H., Qiu, X., and Huang, X. (2020). FLAT: Chinese NER using Flat-lattice transformer. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics *2020*, 6836–6842. https://doi.org/10.18653/v1/2020.acl-main.611.

Liu, P., Qiu, X., and Huang, X. (2016). Recurrent neural network for text classification with multi-task learning. Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence 2016, 2873–2879. https://doi.org/10.48550/arXiv.1605.05101.

Liu, X.F., Xu, X.K., and Wu, Y. (2021). Mobility, exposure, and epidemiological timelines of COVID-19 infections in China outside Hubei province. Sci. Data *8*, 54–57. https://doi.org/10.6084/m9.figshare.13567553.

Malhotra, A., Hepokoski, M., McCowen, K.C., and Y-J Shyy, J. (2020). ACE2, metformin, and COVID-19. iScience *23*, 101425. https://doi.org/10.1016/j.isci.2020.101425.

Moorthy, V., Henao Restrepo, A.M., Preziosi, M.P., and Swaminathan, S. (2020). Data sharing for novel coronavirus (COVID-19). Bull. World Health Organ. *98*, 150. https://doi.org/10.2471/BLT.20.251561.

O'Driscoll, M., Ribeiro Dos Santos, G., Wang, L., Cummings, D.A.T., Azman, A.S., Paireau, J., Fontanet, a., Cauchemez, S., and Salje, H. (2021). Age-specific mortality and immunity patterns of SARS-CoV-2. Nature *590*, 140–145. https://doi.org/10.1038/s41586-020-2918-0.

Salje, H., Tran Kiem, C., Lefrancq, N., Courtejoie, N., Bosetti, P., Paireau, J., Andronico, A., Hozé, N., Richet, J., Dubost, C.-L., et al. (2020). Estimating the burden of SARS-CoV-2 in France. Science *369*, 208–211. https://doi.org/10.1126/science.abc3517.

Singapore Ministry of Health (2020). https://www.moh.gov.sg/news-highlights/details/5-new-cases-of-locally-transmitted-covid-19-infection-31decfullpr.

Sui, D., Chen, Y., Liu, K., Zhao, J., and Liu, S. (2019). Leverage lexical knowledge for Chinese named entity recognition via collaborative graph network. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, *2019*Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3821–3831. https://doi.org/10.18653/v1/D19-1396.

Taiwan. (2020). Ministry of Health and Welfare. https://www.mohw.gov.tw/cp-4632-53100-1.html.

Tian, H., Liu, Y., Li, Y., Wu, C.H., Chen, B., Kraemer, M.U.G., Li, B., Cai, J., Xu, B., Yang, Q., et al. (2020). An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. Science *368*, 638–642. https://doi.org/10.1126/science.abb6105.

United States of America (2021). GlobalHealth. https://www.globalhealth.com.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. Adv. Neural Inf. Process. Syst. *2017*, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762.

Whaiduzzaman, M., Hossain, M.R., Shovon, A.R., Roy, S., Laszka, A., Buyya, R., and Barros, A. (2020). A privacy-preserving mobile and fog computing framework to trace and prevent covid-19 community transmission. IEEE J. Biomed. Health Inform. *24*, 3564–3575. https://doi.org/10.1109/JBHI.2020. 3026060.

World Health Organization (2022). Public Health and Social Measures. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/phsm.

Xu, B., Gutierrez, B., Mekaru, S., Sewalk, K., Goodwin, L., Loskill, A., Cohn, E.L., Hswen, Y., Hill, S.C., Cobo, M.M., et al. (2020a). Epidemiological data from the COVID-19 outbreak, real-time case information. Sci. Data *7*, 1–6. https://doi.org/10.6084/m9.figshare.11974344.

Xu, X.K., Liu, X.F., Wu, Y., Ali, S.T., Du, Z., Bosetti, P., Lau, E.H.Y., Cowling, B.J., and Wang, L. (2020b). Reconstruction of transmission pairs for novel coronavirus disease 2019 (COVID-19) in mainland China: estimation of superspreading events, serial interval, and hazard of infection. Clin. Infect. Dis. *71*, 3163–3167. https://doi.org/10.1093/cid/ciaa790.

Yan, H., Deng, B., Li, X., and Qiu, X. (2019). Tener: adapting transformer encoder for named entity recognition. Preprint at arXiv. arXiv:1911.04474. https://doi.org/10.48550/arXiv.1911.04474.

Yang, H., Sürer, Ö., Duque, D., Morton, D.P., Singh, B., Fox, S.J., Pasco, R., Pierce, K., Rathouz, P., Valencia, V., et al. (2021). Design of COVID-19 staged alert systems to ensure healthcare capacity with minimal closures. Nat. Commun. *12*, 3767. https://doi.org/10.1038/s41467-021-23989-x.

Zhang, J., Litvinova, M., Liang, Y., Wang, Y., Wang, W., Zhao, S., Wu, Q., Merler, S., Viboud, C., Vespignani, A., et al. (2020). Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China. Science *368*, 1481–1486. https://doi.org/10.1126/science.abb8001.

Zhang, Y., and Yang, J. (2018). Chinese NER using Lattice LSTM. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics *2018*, 1554–1564. https://doi.org/10.18653/v1/P18-1144.

Zheng, D., Luo, Q., and Ritchie, B.W. (2021). Afraid to travel after COVID-19? Self-protection, coping and resilience against pandemic 'travel fear. Tourism Manag. *83*, 104261. https://doi.org/10.1016/j.tourman.2020.104261.

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., and Xu, B. (2016). Attention-based bidirectional long short-term memory networks for relation classification. Proceedings of the 54th annual meeting of the association for computational linguistics *2016*, 207–212. https://doi.org/10.18653/v1/P16-2034.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
COVID-19 case reports disclosure in natural language	Liu et al. (Liu et al., 2021)	https://abcdefg3381.github.io/COVID_19_China_case_reports/
Other		
Named Entity Recognition Baseline Model	Lattice	https://github.com/jiesutd/LatticeLSTM
Named Entity Recognition Baseline Model	TENER	https://github.com/fastnlp/TENER
Named Entity Recognition Baseline Model	GraphNER	https://github.com/D2KLab/GraphNER
Named Entity Recognition Baseline Model	FLAT	https://github.com/netless-io/flat
Text Classification Baseline Models	Chinese Text Classification	https://github.com/649453932/Chinese-Text-Classification-Pytorch
Epidemic Record Extraction System	CCIE System	http://covid19.caseassistant.top

RESOURCE AVAILABILITY

Lead contact

Further information and request should be directed to the lead contact, Yuanyuan Sun (syuan@dlut.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Data reported in this paper will be shared by the lead contact upon request. All interested investigators will be allowed access to the COVID case reports once they register and pledge to not re-identify individuals or share the data with a third party. The dataset that contains the annotated fields and categories of case reports is obtainable upon request by contacting the corresponding authors. The code used in this study is listed in key resources table and also available upon reasonable request from authors. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Data preprocessing

We used the natural language case disclosure reports published in the dataset of Liu et al. (https://abcdefg3381.github.io/COVID_19_China_case_reports/) and organized a team of a dozen graduate students who have majored in computational communication or artificial intelligence to manually annotate the case reports. Each case disclosure was encoded into 17 fields, including demographic information, travel history, exposure to known infections, and timelines of case admission. These data fields correspond to two NLP tasks: 11 named-entity recognition tasks and six text classification tasks.

The annotation process has three steps: manual annotation, calibration, and consistency inspection. The manual annotation involves performing field-screening sentence by sentence and determining the field label based on the trigger words or their context. The calibration requires individuals who annotate the same infection case exchange their annotation cases for inspection. Consistency inspection corrects the same infection cases annotated by different individuals by using the machine program that screens for inconsistent field labels and timely feedback to the annotators.

Annotation strategy

Named-entity recognition (Table S4)

We annotated 11 data fields for each case report, namely (1) *age* (AGE), (2) *gender* (GED), (3) *departure place* (SL), (4) *transit place* (TL), (5) *destination place* (DL), (6) *arrival dates* (DT), (7) *quarantine dates* (IT), (8) *symptom onset dates* (OnT), (9) *hospitalization dates* (TT), (10) *confirmation dates* (CT), and (11) *admitted hospital* (TDH). For each named entity, we defined a group of trigger words, i.e., representative words that can clearly indicate the field (e.g., "hospital" is the trigger word for the *admitted hospital* field). We observed that the major difference among the infection cases is the description granularity of the fields, especially those related to location. Take the transit location of an infected person as an example: Some infection cases are accurate to the level of "community," while others are only recorded till the level of "city." Thus, if the text contained multiple expressions belonging to the same data field, they were all labeled under this field. We also added three additional labels (other location, other time, and other institution) in the annotation strategy. Though of little practical use, it is critical to associate the dates and places with vague descriptions to these labels to decrease the possibility of the neural networks recognizing the dates and places as incorrect entities.

We manually coded 1,200 case reports from the data of Liu et al. These samples were chosen by examining the difference between their label distribution and that of the entire dataset. Specifically, a loss function was defined and minimized:

$$Loss = \frac{1}{L}\sum_{i=1}^{L}\left(\left|N^{i}_{gold} - N^{i}_{sample}\right|\right) \quad \text{(Equation 1)}$$

Here, $L = 11$ is the total number of data fields; N^{i}_{gold} and N^{i}_{sample} are the number of the i-th label in the manually coded data and the number of the i-th label in the sampled data, respectively.

Text classification (Table S5)

We annotated six categories for each case report: (i) the location (Place), (ii) event (Event), (iii) individuals (Person) causing possible exposure, (iv) quarantine place (Isolate), (v) methods of detection (Discover), and (vi) degree of clinical symptoms (Degree). We asked the human coders to group the expressions with similar semantics into a predefined set of annotations. Among all categories, the "Event" data field had the largest number of annotations ($n = 8$), whereas "Place" had the least number of annotations ($n = 3$). We adopted text-matching techniques to assign labels to infection cases. We first constructed a vocabulary for each category to capture all possible expressions and the corresponding annotations. Then, we matched all the words in each case report with the vocabulary to determine the most relevant category to which the case should belong. The first 10,000 case reports (i.e., from January 2 to March 4 2020) were annotated.

Structure of CCIE

The CCIE is a two-step framework (Figure 1B). First, the CCIE uses a pretrained language model with the WWM (Cui et al., 2021) to encode case reports to convert each word (token), as well as the entire document, to vector representations. Then, it finetunes the embeddings in downstream tasks. The named-entity-recognition network comprises a Bi-LSTM network and a conditional-random-field (CRF) (Kadari et al., 2017) layer for named-entity-recognition tasks. The text classification network is a fully connected neural network used for text classification tasks.

The *pretrained language model* is a concatenation of a bidirectional transformer (Vaswani et al., 2017). The objective function of this model can be expressed as follows:

$$Objective = P(w_i | w_1, \ldots, w_1, w_{i+1}, w_{i+2,\ldots,} w_n) \quad \text{(Equation 2)}$$

where wi is each word in an infection case report.

The initial input of the model is a set of infection record reports C = {c1, c2,…, cM , where Cm represents the m-th infection case, and $m \in M$ Any infection case c can be represented as c = {w1, w2,…,wN} where wn is the n-th word in the infection case, and $n \in N$. The input vectors $E_i = \{C_{word,} C_{seg,} C_{pos}\} (i \in M)$ of the pretrained language model are the initial vectors of each infection record report ci, comprising the word embedding Cword, segment embedding Cseg, and position embedding Cpos. The pretrained language model uses a 12-layer transformer to learn the contextual information of the words in infection cases. The

core component of the transformer is the multi-head attention mechanism, which can be calculated as follows:

$$Q = E_n * W^Q; K = E_n * W^K; V = E_n * W^V$$

$$MulHead(Q,K,V) = concat(hd_{1,\ldots,}hd_h)W^o \quad \text{(Equation 3)}$$

where $hd_i = Att(QW_i^Q, KW_i^K, VW_i^V)$ and $Att(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \bullet Q, K, V$ are the input embeddings of the attention, representing the query vector, key vector, and value vector, respectively, and d_k represents the dimensions of the input vectors.

The pretrained language model randomly masks 15% words for encoding infection cases, and the objective function calculates only the conditional probability of these 15% masked words. Of all the masked words, 10% are replaced with other words in infection cases, another 10% remain constant, and the remaining 80% are replaced with the **[*mask*]** symbol. In the training phase, the pretrained language model reduces the sentence length to 128 words in 90% of the training periods to improve training efficiency and decrease time consumption. In addition, the pretrained language model can learn the features of long texts by adding the task of predicting the next sentence. Therefore, the vector conversion for infection cases can be summarized as follows:

$$X_n = \text{Pre\_}trained(E_n, \theta) \quad \text{(Equation 4)}$$

where $n \in N, E \in \mathbb{R}^{d_{Pre\_trained}}$ and θ represents the parameters of the pretraining language model. When X_n denotes the real-value embeddings corresponding to each word in infection cases, the output of the pretraining language model is a word vector. When X_n assumes the real-value embeddings of the **[*CLS*]** start symbol, the model's output is the sentence vector.

The *named-entity-recognition network* comprises a Bi-LSTM layer and a CRF layer. It extracts structural information through sequence labeling. It identifies the entity in infection cases based on the word embeddings X_n^{word} obtained from the pretrained language model. Bi-LSTM is a recurrent neural network that can learn the long-distance dependence among entities. The principle of Bi-LSTM is as follows:

$$\begin{aligned}
i_n &= \sigma\left(W_i X_n^{word} + U_i h_{n-1} + b_i\right)\\
f_n &= \sigma\left(W_f X_n^{word} + U_f h_{n-1} + bf\right)\\
\tilde{c}_n &= \tanh\left(W_c X_n^{word} + U_c h_{n-1} + b_c\right)\\
o_n &= \sigma\left(W_o X_n^{word} + U_0 h_{n-1} + b_o\right)\\
c_n &= f_n \odot c_{n-1} + i_n + \odot \tilde{c}_n\\
h_n &= o_n \circ \tanh(c_n)
\end{aligned} \quad \text{(Equation 5)}$$

W and U are two trainable parameters, $n \in N$ and N is the sentence length. The variables $i_n, (f_n, \tilde{c}_n)$ and o_n indicate the input, forget and output gates, respectively. c_n and h_n indicate the cell-state and hidden-state of Bi-LSTM.

Considering the correlation among the entities, CCIE adds a CRF layer behind LSTM, which takes h_n as an input to learn the probability distribution of the entity labels. For a given infection case set $c = \{c_1, c_2, \ldots, c_N\}$, the probability of its label sequence $y = \{l_1, l_{2,\ldots,} l_N\}$ can be calculated as follows:

$$P(y|c) = \frac{\exp\left(\sum_n \left(W_{CRF^{hn}}^{l_n} + b_{CRF}^{(l_{n-1}l_n)}\right)\right)}{\sum_{\hat{y}} \exp\left(\sum_n \left(W_{CRF^{hn}}^{l_n} + b_{CRF}^{(l_{n-1}l_n)}\right)\right)} \quad \text{(Equation 6)}$$

where $y' = \{l'_1, l'_2, \ldots, l'_N\}$ represents any possible label sequence, and $W_{CRF}^{l_n}$ and $b_{CRF}^{(l_{n-1}, l_n)}$ are trainable parameters. Therefore, if there are M training samples $\{(c_i, y_i)\}|_{i=1}^N$, then the loss function of the named-entity-recognition network can be expressed as follows:

$$L = -\sum^N \log(P(y_i|c_i)) \quad \text{(Equation 7)}$$

The *text classification network* comprises a fully connected neural network. It aims to predict the true annotation of the entire case report based on the sentence vector $X_n^{sentence}$ obtained from the pretrained language model. For M given infection cases $S_i|_{i=1}^{M}$ and their annotations $y_i|_{i=1}^{M}$, the loss function of text classification tasks network is calculated as follows:

$$P(y|s) = -\sum_{i=1}^{M} y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \quad \text{(Equation 8)}$$

where i represents the$i$-th report, $\hat{y}$ represents the annotation predicted by CCIE, and y represents the true annotation of the infection case.

Model training

The samples used for model training and validation were collected only from officially released public case reports. All the collected data was anonymized for the purpose of this study. The study protocol was reviewed and approved by the original data publisher.

For the training and evaluation of the CCIE framework, we adopted the traditional evaluation method of deep-neural-network models, which divided the unified dataset into training, verification, and test sets. The training set was used to train the model parameters, and the verification set was used to select the best model. For the best model, the test set was used to evaluate the model's performance. The verification set can be a part of the data separated from the training set, but the test data must never be considered in the training process, which means that these data are completely invisible to the CCIE framework.

To recognize the entities, we used 80% of the annotated data for training, 10% for verification, and the remaining 10% for testing. In the training stage, we set the training period to 32 and the word embedding dimensions to 768. We evaluated the label prediction performance in terms of precision (P), recall (R), and F1 values (F); the formulations for these three evaluations are as follows:

$$P = \frac{TP}{TP + FP}; R + \frac{TP}{TP + FN}; F = \frac{2 \cdot P \cdot R}{P + R} \quad \text{(Equation 9)}$$

where TP indicates the number of correct predictions of positive samples, FP indicates the number of incorrect predictions of positive samples, and FN indicates the number of incorrect predictions of negative samples. The F value is the harmonic mean value of precision and recall. To obtain objective results, the experiment was conducted three times on the dataset, and the results were then averaged to obtain the final result.

For the training of the text classification network, we set the training period to 50 and the sentence embedding dimensions to 768. We employed the weighted F value to evaluate CCIE, and the formulation is as follows:

$$weighted\_F = \frac{1}{n}\sum_{i-1}^{K} F_i \cdot W_i \quad \text{(Equation 10)}$$

where K represents the number of label types, F_i represents the F value of each category i, and W represents the weight matrix (the number of labels in each category is used as the weight).

Parameter setting

The main parameters of our CCIE are as follows: (i) The pretraining model is Roberta-WWM_ext_Large-12-768 containing a 12-layers transformer, where Roberta is trained with the WWM mechanism. (ii) The named-entity-recognition network contains BiLSTM and CRF. BiLSTM employs a two-layer neural network to reduce the word embeddings to 300 dimensions. In the training stage, the number of training periods was 40, and the batch size in each training iteration was 32. (iii) The text classification network is a fully connected layer. In the training stage, the number of training periods was 50, and the batch size was set to 32 in each period.

The main parameters used in baseline models for the named-entity-recognition task are as follows: (i) The LSTM in Lattice (Zhang and Yang, 2018) uses one-layer architecture and 200 hidden units to compute word embeddings. The learning rate is set to 0.015, and the dropout is set to 0.5. (ii) TENER (Yan et al., 2019) employs two blocks of transformer architecture and four heads in the transformer. The training period is set 50, and the batch size is set to 16 in each period. The learning rate is set to 7e-4, and the dropout is

set to 0.15. (iii) GraphNER (Sui et al., 2019) adopts one graph convolution layer to encode case reports. The training period is set to 5, and the batch size is set to 64 in each period. The learning rate is set to 5e-4, and the dropout is set to 0.5. (iv) FLAT (Li et al., 2020) uses one block of transformer architecture and four heads in the transformer. The training period is set to 100, and the batch is set to 10 in each period. The learning rate is set to 6e-4, and the dropout is set to 0.5. These parameters are the optimal settings that are picked from the source code published in the original literature.

The baseline methods used in the text classification task are reproduced from the GitHub library Chinese-Text-Classification-Pytorch (Hu, 2020). Hence, we preserve the same parameters for these benchmarks to conduct experiments. The main parameters are as follows: The dropout is set to 0.5, the padding size is set to 32, the hidden unit is set to 1024, the number of transformer layers is set to 1, the learning rate is set to 5e-4, the dropout is set to 0.5, the training period is set to 20, and the batch size is set to 128 in each period.

GitHub entries for the source codes of the baseline methods are listed in the key resources table.

QUANTIFICATION AND STATISTICAL ANALYSIS

All experiments and evaluations were performed using a Linux system with a GPU (3090), a CPU of 48 cores, and 128 GB of memory. The t-test was performed using the SPSS tool.

ADDITIONAL RESOURCES

The Epidemic Record Extraction System to help extract structured data fields from open-access COVID-19 case reports: http://covid19.caseassistant.top.