Title:

Epidemiologic Information Discovery from Open-access COVID-19 Case Reports via Pretrained Language Model

Authors:

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

Affiliations:

[1] College of Computer Science and Technology, Dalian University of Technology, Liaoning, 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, New Mexico, 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.

[†] These authors contributed equally to this paper: Zhizheng Wang, Xiaofan Liu, Zhanwei Du, Lin Wang.

[*] Corresponding authors: Zoie Shui-Yee Wong (zoiesywong@gmail.com), Xiao-Ke Xu (xuxiaoke@foxmail.com), Yuanyuan Sun (syuan@dlut.edu.cn).

HIGHLIGHTS

- We propose a computational framework that can automatically extract epidemiological information from open-access COVID-19 case reports.
- The information extracted from our approach is highly consistent with that obtained from the gold-standard manual coding, with a matching rate of 80%.
- We provide an open-access online platform that can accurately estimate epidemiological statistics in real-time with substantially reduced burden in data curation.

ABSTRACT (149 words)

Although open-access data are increasingly common and useful to epidemiological research, curation of such datasets is resource-intensive and time-consuming. Despite a major source of COVID-19 data, the regularly disclosed case reports were often written in natural language with 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 language model developed using deep neural networks with training samples compiled using an optimized data annotation strategy. When applying to the COVID-19 case reports collected from mainland China, our novel framework outstrips 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 lower burden in data curation.

Keywords: Epidemiologic information; Artificial Intelligence; COVID-19 case reports; Pretrained language model; Epidemiology; Natural language processing.

INTRODUCTION

The coronavirus disease 2019 (COVID-19) pandemic has been a global public health crisis [1],[2],[3],[4], with more than 300 million confirmed cases as of the end of 2021. Many countries and regions, such as China [5], Singapore [6], and Taiwan [7] were able to publish COVID-19 case reports obtained from detailed epidemiological investigations in real-time, with the aim to enhance situation awareness [8] and promote individual behavior of self-protection [9]. 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 [10][11][12][13][14][15], to estimate the infection and disease burdens [16][17], characterize population behavioral changes [18][19],[20], and optimize control measures [21][22][23]. 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 distinct words to express the same information). The data curation and standardization processes can be resource-intensive and time-consuming [24]. For example, Liu *et al.* [5] recruited a team of twenty data curators to trace and manually annotate the demographic information, mobility history, and epidemiological timelines for each COVID-19 case that publicly reported from mainland China as of 4 March 2020. To reduce the burden on human resources and facilitate the real-time analyses of open-access case reports, the early research [25] inspires us to develop a deep learning framework using natural language processing (NLP) techniques to automatically identify key information from the raw case reports (**Fig. 1a**).

Generally, different from the rule-based methods that use regularization formulation to match line lists from raw data, a deep learning framework curating the open case reports involves a combination of named entity recognition, text classification, and knowledge inference tasks in NLP. For example, the identification of spatial location and calendar dates requires named entity recognition. Distinguishing the case detection method such as the reverse transcription polymerase chain reaction (RT-PCR) test or symptom onset requires text classification, as the expressions often vary with different natural language writing styles in the reports. Where there were incomplete data fields or vague language expressions, these require knowledge inference and standardization (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 pre-trained language model [26] directly to the human-coded data (**Fig. S1**).

Therefore, we require adjusting the existing NLP models for our data curation task via preparing appropriately high-quality annotated data. We first propose a machine-learnable annotation strategy to refine the codebook in ref. [5], 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 that uses three deep neural networks to perform the named entity recognition and text classification (**Fig. 1b**). The pre-trained language model with the whole-word mark (WWM) mechanism [27] encodes each case report into vector representations, a bidirectional-long short-term memory (Bi-LSTM) [28] performs the named entity recognition and a fully connected network performs the text classification. At last, we evaluate CCIE in three aspects. Firstly, we apply our annotation strategy to different deep neural networks to observe the adaptability of annotated data. Secondly, we compare the proposed CCIE framework with state-of-the-art models in the task of named entity recognition and text classification. Finally, we investigate the effectiveness of the CCIE framework through cross-validation with manually extracted values. In Practice, we also develop an online system based on CCIE publicly available for all researchers worldwide.

RESULTS

Performance evaluation

***Annotation strategy*.** The annotation of case reports is used as labels for different deep neural networks. To guarantee the consistency and accuracy of manual annotation, we randomly examined and modified a subset of 100 case reports after the annotation 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 consensus on the modifications. Then, we continued to examine and manually annotate the remaining case reports. The agreement rate on our revisions reaches 90%, which suggests that the inter-annotator agreement rate is acceptable. Any inconsistent revisions were provided to the experts for final revisions. Based on our annotation data, all deep neural networks used in this study demonstrate high adaptability across different tasks (i.e., named entity recognition and text classification). For example, in the named entity recognition tasks (**Fig. 2a**), all models achieve F1-values higher than 70% for most entities. For fields with a fixed language format, such "dates", and obvious trigger words, such as "admitted hospital", the F1-value (a global evaluation for precision and recall, which is calculated by Eq. (8)) of 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 obtain F1-values higher than 50%. In the text classification tasks (**Fig. 2b**), most models achieve F1-values higher than 80% in the data fields with limited labels. Even for the category with more possible labels, such as "event", the evaluated models obtain an F1-value exceeding 75%.

Text classification tasks*.** To further analyze the performance of CCIE, we first compare it with seven benchmark text classification algorithms (i.e., Transformer [29], DPCNN [30], FastText [31], TextCNN [32], TextRNN [33], TextRCNN [34], and LSTM [35]) in the classification of six categories (Tab. S1(a)***). The F1-values obtained by CCIE in all six categories are above 82%, with the highest value reaching 93.2%. Especially in the *event* category, with maximum category labels, CCIE increases by 2.6% compared to TextCNN and by 10.7% compared to Transformer. This result shows that the pretrained language models obtain word embeddings with richer semantic expression for mining deep features in the text, such as syntactic dependence and semantic role.

Named entity recognition tasks**.** Then, we also compare CCIE with four classical deep neural network models (i.e., Lattice [36], TENER [37], GraphNER [38], and FLAT [39]) in the recognition of nine entities (***Tab. S1(b)***). CCIE demonstrates the best performance in most entity recognitions (7/9 cases), including all "dates" and two "places". This achievement is attributed to the fact that the pretrained model used in CCIE adopts the whole-word mark (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* "destination place", as the description granularities of these fields are recorded as "xx City" and "xx County," respectively. In addition, we compare CCIE with other pretrained language model-based solutions (i.e., "TENER + BERT" and "TENER + (BERT with WWM)"). The results show that CCIE outperforms "TENER + BERT" and obtains comparable results with "TENER + (BERT with WWM)" (***Tab. S1(c)***), 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 conduct the named entity recognition task with CCIE under different annotated data volumes. With an increase in the annotated data size, the overall performance of named entity recognition shows an upward trend (**Fig. 3**). However, when the annotated data size reaches 400, the upward trend is no longer evident, and the revenue curve appears to be stable between

0.1% and 0.2%. Moreover, from the perspective of the recognition accuracy of each entity (***Fig. S2***), the result and revenue stop fluctuating significantly when the annotated data size reaches 600, and the average revenue remains between 0% and 0.04%. Note that the recognition of *admitted hospital* reaches an optimum value (90%) when the annotated data size is 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 generates large differences in their language styles. Moreover, our CCIE is a feature learning model and, therefore, sensitive to language styles. We select eight provinces that report the highest number of cases (i.e., Zhejiang, Jiangsu, Shandong, Guangdong, Chongqing, Hunan, Anhui, and Henan) and compare the CCIE performance for the reports released by each province (**Fig. 4a**). CCIE performs well on the reports released by the health departments in Zhejiang (91.67%), Jiangsu (89.33%), and Shandong (88.23%) but not on those released by the health departments in Guangdong (78.82%) and Chongqing (76.28%).

After parsing the report examples released by different provinces (**Fig. 4b**), we found that case reports can be most easily processed by 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 (**Tab.S2**) and design the corresponding questions that should be asked in epidemiology surveys (***Tab. S3***), which covers the travel history and the social (contact) behaviors.

Cross-validation with manually extracted information

We compare the 17 machine-extracted fields with the manually extracted ones in Liu et al.'s dataset [5] on the first 10,000 case disclosure reports (i.e., from January 2 to March 4, 2020). We use a simple fuzzy matching logic (**Fig. 5a**) to deal with the style differences between the machine- and human- extracted fields—if the machine-extracted text is present in the manually coded fields, we consider that the machine has provided meaningful information. The comparison results (**Fig. 5b**) show high consistency between machine extraction and manual coding. The agreements of *ages* and *genders* are 97.2% and 97.96%; those of *the places of departure*, *transit*, and *destination* are 74.95%, 86.84%, and 66.62%; and those of the *dates of arrival*, *quarantine*, *symptom onset*, *hospitalization*, and *confirmation* are 87.67%, 75.14%, 86.21%, 69.24%, and 65.89%, respectively.

The matching rate for the *admitted hospital* field is relatively low. We find that the inconsistency between machine extraction and human coding lies in the following facts: (i) the machine extracts the abbreviations of hospital names, while human coding converts them to full names (~85% of the cases); (ii) a vague term "designated medical institution" used by the local authorities instead of the exact hospital names is not considered in human coding but recognized by the machine (~10% of the cases); and (iii) the machine fails to recognize the correct *admitted hospital* field mentioned in the report or does not recognize it at all (~5% of the cases).

The machine-extracted information can also be used to determine the epidemiological characteristics of COVID-19 with close-to-human coding precision. For example, we use *the date of the symptom onset* field to compute the real-time reproduction (RT) number. During January 8 to February 26, 2020, the distribution of RT values calculated by human coding and machine extraction remains consistent. We also calculate the R-square value and rooted mean squared error value of these two groups of numerical distributions (**Fig. 6**). The results demonstrate a high consistency between two distributions, which also shows that the field extracted by the machine has a comparable result with human encoding in the calculation of the RT 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 organize the location fields extracted from an infection case into a timeline based on temporal logic, so that researchers can more intuitively grasp the activity trajectory of infected patients. We also add 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 on the GPU is five seconds per case. The average processing speed on the CPU is approximately ten seconds per case.

DISCUSSIONS

The epidemiological analysis of community transmission is vital for formulating public health interventions against COVID-19 [40][41]. This is critical for clarifying the host selection and physiological mechanism of COVID-19 by obtaining essential contents, such as the gathering behavior and activity trajectory from the massive infection cases. To facilitate the automatic extract of epidemiological information from open-access COVID-19 case reports, we first propose a refined annotation strategy based on available human coding and then develop 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 outstrips the performance of several state-of-the-art models such as LSTM [35], Transformer [29], Lattice [36], TENER [37]. In particular, our method on average reduces around 80% of the labor (about twenty annotators) who works on the manual coding of raw case reports written in natural language, and the machine-extracted data fields are able to correct some incorrectly coded field by humans such as the inconsistency in the word segments extracted for *admitted hospital*.

To ease the implementation of our framework, we provide an online system that can be access through this website: *http://covid19.caseassistant.top*. This system allows users to extract all 17 data fields analyzed in our study from their own case reports. It will serve as a preliminary step in the automatic information extraction of epidemiology survey reports and is expected to benefit the wider research community.

Our system that automatically extracts key epidemiological information including demographic, travel history, contact scenarios, and epidemiology timeline information from the open-access case reports has a great potential to accelerate the COVID-19 research. Although we only focus on the case reports written in Chinese language here, our CCIE framework can be easily adapted for applying to other languages. This is because our annotation strategy can be applied to case reports written in different language styles, and we can easily change the pre-trained language model used for Chinese language to the most suitable models for another 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, the 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 [42].

Therefore, we call for standardization of the case reporting format and propose additional questions that should be asked in epidemiology surveys (***Tab. 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 [43], 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 [44]. Although there are exemplar communities [45] hosting data repositories, the lack of structure hinders data processing at a large scale [46]. 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 the automated data extraction, which can be easily extended to data structured processing attributed to the flexibility of neural network models, of publicly available unstructured data. 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.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China [61773091 and 62173065 to X.-K. X., 11975025 to L. W., 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.]; and Japan Society for the Promotion of Science KAKENHI [Grant No. 18H03336 to Z. S.Y. W.].

AUTHOR CONTRIBUTIONS

Methodology: Xiao-Ke Xu, Yuanyuan Sun; Investigation: Zhizheng Wang, Xiao Fan Liu, Zhanwei Du, Lin Wang; Visualization: Zhizheng Wang, Zhanwei Du; Funding acquisition: Xiao-Ke Xu, Lin Wang, Xiao Fan Liu; Supervision: Ye Wu, Petter Holme, Michael Lachmann, Hongfei Lin, Zoie S.Y. Wong; Writing – original draft: Zhizheng Wang, Xiao Fan Liu, Zhanwei Du, Lin Wang; Writing – review & editing: Zhizheng Wang, Xiao Fan Liu, Zhanwei Du, Zoie S.Y. Wong, Xiao-Ke Xu, Yuanyuan Sun

CONFLICT OF INTEREST

All authors declare no competing interests.

REFERENCE

[1]. Malhotra, A., Hepokoski, M., McCowen, K. C., & Shyy, J. Y. ACE2, metformin, and COVID-19. iScience, 23(9), 101425 (2020).

[2]. Andreadakis Z, Kumar A, Román R G, et al. The COVID-19 vaccine development landscape[J]. Nature reviews. Drug discovery, 2020, 19(5): 305-306.

[3]. Agbehadji I E, Awuzie B O, Ngowi A B, et al. Review of big data analytics, artificial intelligence and nature-inspired computing models towards accurate detection of COVID-19 pandemic cases and contact tracing[J]. International journal of environmental research and public health, 2020, 17(15): 5330.

[4]. Chinazzi M, Davis J T, Ajelli M, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak[J]. Science, 2020, 368(6489): 395-400.

[5]. Liu X F, Xu X K, Wu Y. Mobility, exposure, and epidemiological timelines of COVID-19 infections in China outside Hubei province[J]. Scientific data, 2021, 8(1): 1-7.

[6]. MINISTRY OF HEALTH, https://www.moh.gov.sg/news-highlights/details/5-new-cases-of-locally-transmitted-covid-19-infection-31decfullpr. [DB/CD].

[7]. Ministry of Health and Welfare, https://www.mohw.gov.tw/cp-4632-53100-1.html. [DB/CD].

[8]. Deborah Bunker. Who do you trust? The digital destruction of shared situational awareness and the COVID-19 infodemic[J]. International Journal of Information Management, 2020, 55:102201.

[9]. Danni Zheng, Qiuju Luo, Brent W. Ritchie. Afraid to travel after COVID-19? Self-protection, coping and resilience against pandemic ‘travel fear’[J]. Tourism Management, 2021, 83:104261.

[10]. GlobalHealth, https://www.globalhealth.com. [DB/CD].

[11]. GISAID, https://www.gisaid.org/. [DB/CD].

[12]. Xu B, Gutierrez B, Mekaru S, et al. Epidemiological data from the COVID-19 outbreak, real-time case information[J]. Scientific data, 2020, 7(1): 1-6.

[13]. Zhanwei Du, Xiao-Ke Xu, Lin Wang, Spencer J. Fox, Benjamin J. Cowling, Alison P. Galvani, and Lauren Ancel Meyers*. Effects of proactive social distancing on COVID-19 outbreaks in 58 cities, China. Emerging Infectious Diseases, 2020, 26(9): 2269-2271.

[14]. Sheikh Taslim Ali#, Lin Wang#, Eric HY Lau#, Xiao-Ke Xu, Zhanwei Du, Ye Wu, Gabriel M. Leung, Benjamin J. Cowling*, Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions, Science, 2020, 369: 1106-1109.

[15]. Xiao-Ke Xu#, Xiao Fan Liu#, Ye Wu#, Sheikh Taslim Ali#, Zhanwei Du#, Paolo Bosetti, Eric H Y Lau, Benjamin J Cowling, Lin Wang*, Reconstruction of Transmission Pairs for novel Coronavirus Disease 2019 (COVID-19) in mainland China: Estimation of Super-spreading Events, Serial Interval, and Hazard of Infection, Clinical Infectious Diseases, 2020, 71(12):3163-3167.

[16]. O'Driscoll M, Dos Santos G R, Wang L, et al. Age-specific mortality and immunity patterns of SARS-CoV-2[J]. Nature, 2021, 590(7844): 140-145.

[17]. Salje H, Kiem C T, Lefrancq N, et al. Estimating the burden of SARS-CoV-2 in France[J]. Science, 2020, 369(6500): 208-211.

[18]. Zhang J, Litvinova M, Liang Y, et al. Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China[J]. Science, 2020, 368(6498): 1481-1486.

[19]. Du Z, Wang L, Cauchemez S, et al. Risk for transportation of coronavirus disease from Wuhan to other cities in China[J]. Emerging infectious diseases, 2020, 26(5): 1049.

[20]. Tian H, Liu Y, Li Y, et al. An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China[J]. Science, 2020, 368(6491): 638-642.

[21]. Hale T, Angrist N, Goldszmidt R, et al. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker) [J]. Nature Human Behaviour, 2021, 5(4): 529-538.

[22]. Yang H, Sürer Ö, Duque D, et al. Design of COVID-19 staged alert systems

to ensure healthcare capacity with minimal closures[J]. Nature Communications, 2021, 12(1): 1-7.

[23]. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/phsm [DB/CD].

[24]. Kraemer, M. U. G., Scarpino, S. V., Marivate, V., Gutierrez, B., Xu, B., Lee, G., Brownstein, J. S. Data curation during a pandemic and lessons learned from COVID-19. *Nature Computational Science,* 1, 9-10 (2021).

[25]. Saurav Ghosh, Prithwish Chakraborty, Bryan L. Lewis, Maimuna S. Majumder, Emily Cohn, John S. Brownstein, Madhav V. Marathe, and Naren Ramakrishnan. 2017. GELL: Automatic Extraction of Epidemiological Line Lists from Open Sources. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17). Association for Computing Machinery, New York, NY, USA, 1477–1485.

[26]. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[27]. Y. Cui, W. Che, T. Liu, B. Qin and Z. Yang, "Pre-Training With Whole Word Masking for Chinese BERT," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29 : 3504-3514.

[28]. Rekia Kadari, Yu Zhang, Weinan Zhang, Ting Liu. CCG supertagging via Bidirectional LSTM-CRF neural architecture [J]. Neurocomputing, 2018, 283: 31-37.

[29]. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

[30]. Johnson R, Zhang T. Deep pyramid convolutional neural networks for text categorization[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 562-570.

[31]. Joulin A, Grave E, Bojanowski P, et al. Bag of tricks for efficient text classification[J]. arXiv preprint arXiv:1607.01759, 2016.

[32]. Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

[33]. Liu P, Qiu X, Huang X. Recurrent neural network for text classification with multi-task learning[C]//Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. 2016: 2873-2879.

[34]. Lai S, Xu L, Liu K, et al. Recurrent convolutional neural networks for text classification[C]//Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015, 2267-2273.

[35]. Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers). 2016: 207-212.

[36]. Zhang Y, Yang J. Chinese NER Using Lattice LSTM[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 1554-1564.

[37]. Yan H, Deng B, Li X, et al. Tener: Adapting transformer encoder for named entity recognition[J]. arXiv preprint arXiv:1911.04474, 2019.

[38]. Sui D, Chen Y, Liu K, et al. Leverage lexical knowledge for Chinese named entity recognition via collaborative graph network[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 3821-3831.

[39]. Li X, Yan H, Qiu X, et al. FLAT: Chinese NER Using Flat-Lattice Transformer[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 6836-6842.

[40]. Byambasuren O, Cardona M, Bell K, et al. Estimating the extent of

asymptomatic COVID-19 and its potential for community transmission: systematic review and meta-analysis[J]. Official Journal of the Association of Medical Microbiology and Infectious Disease Canada, 2020, 5(4): 223-234.

[41]. Whaiduzzaman M, Hossain M R, Shovon A R, et al. A privacy-preserving mobile and fog computing framework to trace and prevent covid-19 community transmission[J]. IEEE Journal of Biomedical and Health Informatics, 2020, 24(12): 3564-3575.

[42]. http://brat.nlplab.org/manual.html. [DB/CD].

[43]. Beijing Preventive Medicine Association. Guideline for epidemiological investigation of coronavirus disease 2019 (T/BPMA 0003-2020) [J]. Zhonghua liu xing bing xue za zhi, 2020, 41(8): 1184-1191.

[44]. Moorthy, V., Restrepo, A. M. H., Preziosi, M. P., & Swaminathan, S. (2020). Data sharing for novel coronavirus (COVID-19). Bulletin of the World Health Organization, 98(3), 150.

[45]. Global.health: a Data Science Initiative. https://global.health/. [DB/CD].

[46]. Gardner, L., Ratcliff, J., Dong, E., & Katz, A. A need for open public data standards and sharing in light of COVID-19. The Lancet Infectious Diseases, 2021, 21(4), e80.

[47]. Cori A, Ferguson N M, Fraser C, et al. A new framework and software to estimate time-varying reproduction numbers during epidemics[J]. American journal of epidemiology, 2013, 178(9): 1505-1512.

STAR★METHOD

KEY RESOURCE TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
COVID-19 case reports disclosure in natural language	Liu et al. [5]	https://abcdefg3381.github.io/COVID_19_China_case_reports/
Software and algorithms		
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, Zoie Shui-Yee Wong (zoiesywong@gmail.com), Xiao-Ke Xu (xuxiaoke@foxmail.com), Yuanyuan Sun (syuan@dlut.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

All interested investigators will be allowed access to the COVID case reports following registration and pledging to not re-identify individuals or share the data with a third party. The data set that contains annotated fields and categories of case reports is obtainable upon request by contacting the corresponding authors. The code used in this study is available upon reasonable request from authors.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Our study does not use experimental models typical in life sciences.

METHOD DETAILS

Data preprocessing

We use the natural language case disclosure reports published in Liu et al.'s dataset (*https://abcdefg3381.github.io/COVID_19_China_case_reports/*) and organize a team of a dozen graduate students who major in computational communication or artificial intelligence to manually annotate the case reports. Each case disclosure is 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: eleven named entity recognition tasks, and six text classification tasks.

The annotation process has undergone three steps: manual annotation, calibration, and consistency inspection. The manual annotation needs to perform field screening sentence by sentence and determine the field label based on the trigger words or their context. The calibration requires that the annotation personnel who annotates the same infection case exchange their document for inspection. The consistency inspection uses the machine program to correct the same infection cases annotated by different personnel, which screens for inconsistent field labels and timely feedback to the annotation personnel.

Annotation strategy

Named entity recognition (***Tab. S4***). We annotate 11 data fields for each case report, including (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 define 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 observe that the major difference among the infection cases is the description granularity of 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 other infection cases are only recorded to the level of “city.” Therefore, if the text contains multiple expressions belonging to the same data fields, they are all labeled under this field. We also add three additional labels (i.e., 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 for decreasing the possibility that the neural networks recognize the dates and places as incorrect entities.

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

$$Loss = \frac{1}{L}\sum_{i=1}^{L}(|N_{\text{gold}}^{i} - N_{\text{sample}}^{i}|) \qquad (1)$$

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

Text classification (***Tab. S5***)**.** We annotate 6 categories for each case report, including (i) to (iii) the location (Place), event (Event), and individuals (Person) causing possible exposure, (iv) quarantine place (Isolate), (v) methods of detection (Discover), and (vi) degree of clinical symptoms (Degree). We ask the human coders to group the expressions with similar semantics into a predefined set of annotations. Among all categories, the “Event” data field has the largest number of annotations (*n* = 8), whereas the “Place” data field has the least number of annotations (*n* = 3). We adopt text matching technique to assign labels to infection cases. We first construct a vocabulary for each category to store all possible expressions and the corresponding annotations. Then, we match all words in each case report with the vocabulary to determine the most relevant annotation to which the case should belong. The first 10,000 case reports (i.e., from January 20 to March 4, 2020) are annotated.

Structure of CCIE

CCIE is a two-step framework (**Fig. 1(b)**). First, CCIE uses a pretrained language model with the whole-word mark (WWM) [27] to encode case reports to convert each word (token), as well as the entire document, to vector representations. Then, it fine-tunes the embeddings in downstream tasks. The named entity recognition network comprises a Bi-LSTM network and a conditional random field (CRF) [28] layer for named entity recognition tasks. The text classification network is a fully connected neural network for text classification tasks.

Pretrained language model is a concatenation of a bidirectional transformer [29]. The objective function of this model can be expressed as follows:

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

where w_i is each word in an infection case report.

The initial input of the model is a set of infection record report $C = \{c_1, c_2, \cdots, c_M\}$, where c_m represents the m-th infection case and $m \in M.$ Any infection case c can be represented as $c = \{w_1, w_2, \cdots, w_N\}$, where w_n is the n-th word in the infection case and $n \in N$. The input vectors $\boldsymbol{E}_i = \left\{C_{\text{word}},\ C_{\text{seg}},\ C_{\text{pos}}\right\}$ $(i \in M)$ of the pretrained language model are the initial vectors of each infection record report c_i, comprising the word embedding C_{word}, segment embedding C_{seg}, and position embedding C_{pos}.

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

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

$$\text{MulHead}(Q, K, V) = \text{concat}(hd_1, \dots, hd_h)W^O \tag{2}$$

where $hd_i = \mathrm{Att}(QW_i^Q, KW_i^K, VW_i^V)$ and $\mathrm{Att}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$. 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. Among all masked words, 10% are replaced with other words in infection cases, the other 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 in 90% of training epochs to improve the training efficiency and decrease the time consumption. In addition, the pretrained language model can learn the features for long texts by adding the task of predicting the next sentence. Therefore, the vector conversion for infection cases can be summarized as follows:

$$X_n = \mathrm{Pre\_trained}(\boldsymbol{E}_n, \theta) \quad (3)$$

where $n \in N$, $\boldsymbol{E} \in \mathbb{R}^{d_{\mathrm{Pre\_trained}}}$, and θ represents the parameters of the pretraining language model. When X_n takes 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 takes the real-value embeddings of the **[*CLS*]** start symbol, the model's output is the sentence vector.

Named entity recognition network comprises a Bi-LSTM layer and a CRF layer. It extracts structural information by 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 long-distance dependence among entities. The principle of Bi-LSTM is as follows:

$$f_n = \sigma(\boldsymbol{W}_f X_n^{\mathrm{word}} + \boldsymbol{U}_f h_{n-1} + b_f)$$

$$i_n = \sigma\left(\boldsymbol{W}_i X_n^{\mathrm{word}} + \boldsymbol{U}_i h_{n-1} + b_i\right)$$

$$o_n = \sigma\left(\boldsymbol{W}_o X_n^{\mathrm{word}} + \boldsymbol{U}_o h_{n-1} + b_o\right)$$

$$\tilde{c}_n = tanh\left(\boldsymbol{W}_c X_n^{\text{word}} + \boldsymbol{U}_c h_{n-1} + b_c\right)$$

$$c_n \ = \ f_n \ \odot \ c_{n-1} \ + \ i_n \ \odot \ \tilde{c}_n$$

$$h_n = o_n \circ tanh(c_n) \tag{4}$$

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

Considering the correlation among the entities, CCIE adds a CRF layer behind LSTM, which takes h_n as the 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 $y = \{l_1, l_2, \cdots, l_N\}$ can be calculated as

$$P(y|c) = \frac{exp\left(\sum_n\left(\boldsymbol{W}_{\text{CRF}}^{l_n} h_n + b_{\text{CRF}}^{(l_{n-1}, l_n)}\right)\right)}{\sum_{y'} exp\left(\sum_n\left(\boldsymbol{W}_{\text{CRF}}^{l_n'} h_n + b_{\text{CRF}}^{(l_{n-1}', l_n')}\right)\right)} \tag{5}$$

where $y' = \{l_1', l_2', \cdots, l_N'\}$ represents any possible label sequence, and $\boldsymbol{W}_{\text{CRF}}^{l_n}$ and $b_{\text{CRF}}^{(l_{n-1}, l_n)}$ 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

$$L = -\sum_{i=1}^{N} \log\left(P(y_i|c_i)\right) \tag{6}$$

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 pretraining 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

$$P(y|s) = -\sum_{i=1}^{M} y^{(i)} \log\hat{y}^{(i)} + (1 - y^{(i)})\log(1 - \hat{y}^{(i)}) \tag{7}$$

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

Samples 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.

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

To recognize the entities, we use 80% of annotated data for training, 10% for verification, and the remaining 10% for testing. In the training stage, we set the training epoch to 32 and the word embedding dimensions to 768. We evaluate the label prediction performance with Precision (P), Recall (R), and F1-value (F); the formulations for these three evaluations are as follows:

$$P = \frac{TP}{TP+FP} \text{ ; R} = \frac{TP}{TP+FN} \text{ ; } F = \frac{2 \cdot P \cdot R}{P+R} \tag{8}$$

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

In the training of text classification, we set the training epoch to 50 and the sentence embedding dimensions to 768. We employ the weighted-F value to evaluate CCIE and the formulation is as follows:

$$\text{weighted\_}F = \frac{1}{n}\sum_{i=1}^{K} F_i \cdot W_i \tag{9}$$

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

Parameter Setting

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

The main parameters used in baseline models for the named entity recognition task are as follows: (i) The LSTM in Lattice uses 1 layer and 200 hidden to fathom word embeddings. The learning rate is set to 0.015 and the dropout is set to 0.5. (ii) TENER employs 2 transform layers and 4 heads in transformer. The training epoch is set 50, and the batch size is set to 16 in each epoch. The learning rate is set to 7e-4 and the dropout is set to 0.15. (iii) GraphNER adopts 1 graph convolution layer to encode case reports. The training epoch is set to 5 and the batch size is set to 64 in each epoch. The learning rate is set to 5e-4 and the dropout is set to 0.5. (iv) FLAT uses one-layer transformer and 4 heads in transformer. The training epoch is set to 100 and the batch is set to 10 in each epoch. The learning rate is set to 6e-4 and the dropout is set to 0.5.

Baseline methods used in the text classification task are reproduced from the GitHub library, Chinese-Text-Classification-Pytorch. Therefore, we set 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 epoch is set to 20 and the batch size is set to 128 in each epoch.

For one can easily access these baseline methods, we provide the GitHub source in the KEY RESOURCE TABLE.

QUANTIFICATION AND STATISTICAL ANALYSIS

All experiments and evaluations are performed under the condition of a Linux system with a GPU (3090), a CPU of 48 cores, and 128 G memory.

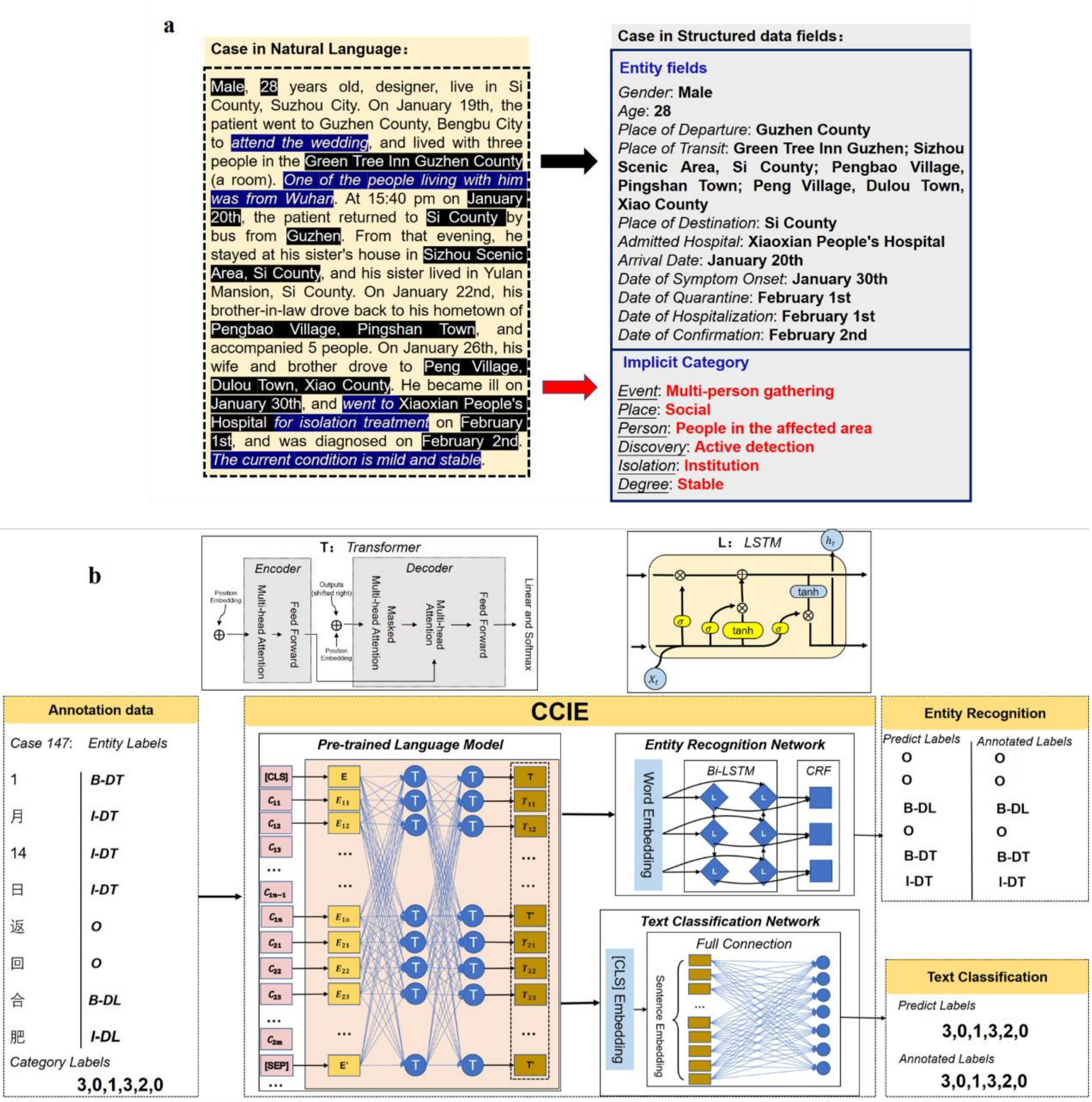

a
Case in Natural Language：
Male, 28 years old, designer, live in Si County, Suzhou City. On January 19th, the patient went to Guzhen County, Bengbu City to attend the wedding, and lived with three people in the Green Tree Inn Guzhen County (a room). One of the people living with him was from Wuhan. At 15:40 pm on January 20th, the patient returned to Si County by bus from Guzhen. From that evening, he stayed at his sister's house in Sizhou Scenic Area, Si County, and his sister lived in Yulan Mansion, Si County. On January 22nd, his brother-in-law drove back to his hometown of Pengbao Village, Pingshan Town, and accompanied 5 people. On January 26th, his wife and brother drove to Peng Village, Dulou Town, Xiao County. He became ill on January 30th, and went to Xiaoxian People's Hospital for isolation treatment on February 1st, and was diagnosed on February 2nd. The current condition is mild and stable.
Case in Structured data fields：
Entity fields
Gender: Male
Age: 28
Place of Departure: Guzhen County
Place of Transit: Green Tree Inn Guzhen; Sizhou Scenic Area, Si County; Pengbao Village, Pingshan Town; Peng Village, Dulou Town, Xiao County
Place of Destination: Si County
Admitted Hospital: Xiaoxian People's Hospital
Arrival Date: January 20th
Date of Symptom Onset: January 30th
Date of Quarantine: February 1st
Date of Hospitalization: February 1st
Date of Confirmation: February 2nd
Implicit Category
Event: Multi-person gathering
Place: Social
Person: People in the affected area
Discovery: Active detection
Isolation: Institution
Degree: Stable
b
T: Transformer
Encoder
Decoder
Position Embedding
Multi-head Attention
Feed Forward
Outputs (shifted right)
Position Embedding
Masked Multi-head Attention
Multi-head Attention
Feed Forward
Linear and Softmax
L: LSTM
h_t
X_t
σ
tanh
Annotation data
Case 147: Entity Labels
1 B-DT
月 I-DT
14 I-DT
日 I-DT
返 O
回 O
合 B-DL
肥 I-DL
Category Labels
3,0,1,3,2,0
CCIE
Pre-trained Language Model
[CLS]
[SEP]
E
E'
T
T'
Entity Recognition Network
Word Embedding
Bi-LSTM
CRF
L
Text Classification Network
[CLS] Embedding
Full Connection
Sentence Embedding
Entity Recognition
Predict Labels
Annotated Labels
O O
O O
B-DL B-DL
O O
B-DT B-DT
I-DT I-DT
Text Classification
Predict Labels
3,0,1,3,2,0
Annotated Labels
3,0,1,3,2,0

Figure 1. COVID-19 cases information extraction (CCIE) framework. **a.** The CCIE 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 contains entity labels and category labels, with the letter "B-" indicating the start position of an entity, and "I-" the middle or end position of an entity. The CCIE comprises a pre-trained language model, a named entity recognition network, and a text classification network. Specifically, the pre-trained 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] the separator between two adjacent sentences. The panel "T: Transformer" explains the internal structure of *Transformer*,

with symbol $\oplus$ indicating the concatenation operation. The named entity recognition network is built with the bidirectional long short-term memory (BiLSTM) model and conditional random fields (CRF) predictor. The panel "L: LSTM" explains the internal structure of this named entity recognition network, with symbol $\otimes$ indicating the elementwise multiplication, σ the sigmoid function, $tanh$ the activation function, and X_t and h_t the input and output of $BiLSTM$, respectively. The text classification network contains a fully connected neural network, with [CLS] vector denoting the sentence embedding. The assessment of model performance requires evaluating both the named entity recognition and text classification.

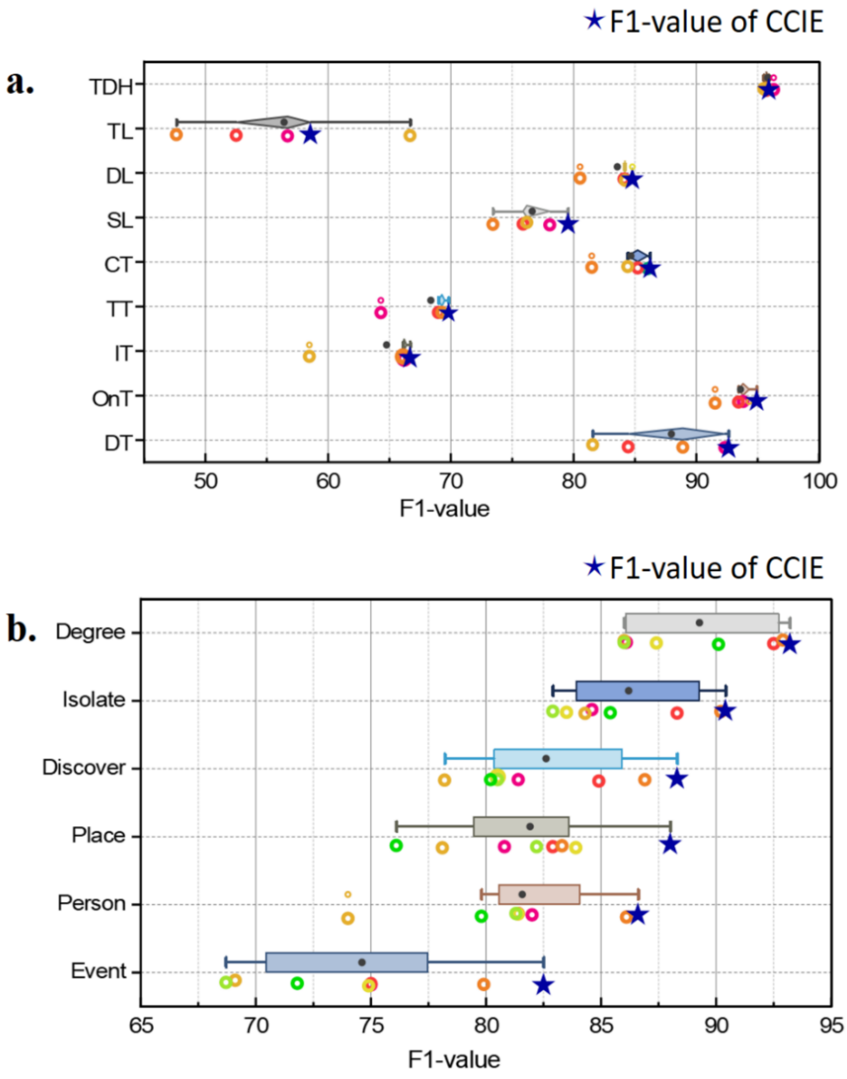


Figure 2. Distributions of F1-value for the named entity recognition and text classification models, using our proposed annotation strategy. (**a**) Distribution of F1-value for each named entity, which aggregates the results of five different named entity recognition methods including Lattice [19], TENER [20], GraphNER [21], FLAT [22], and our CCIE (denoted as "★"). The colored distributions correspond to different named entities. (**b**) Distribution of F1-value for each text category, which aggregates the results of eight text classification methods including DPCNN [24], FastText [25], TextCNN [26], TextRNN [27], TextRCNN [28], LSTM [29], Transformer [30], and our CCIE (denoted as "★"). In each named entity or category, scattered dots indicate the F1-values obtained from different methods, which are used to fit the distribution as indicated by the box plot.

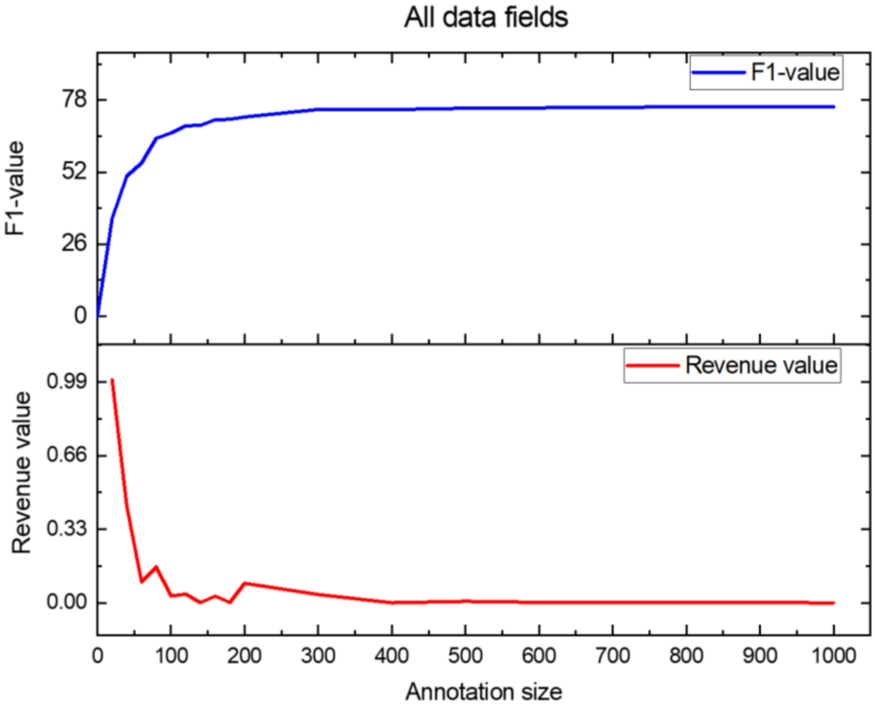
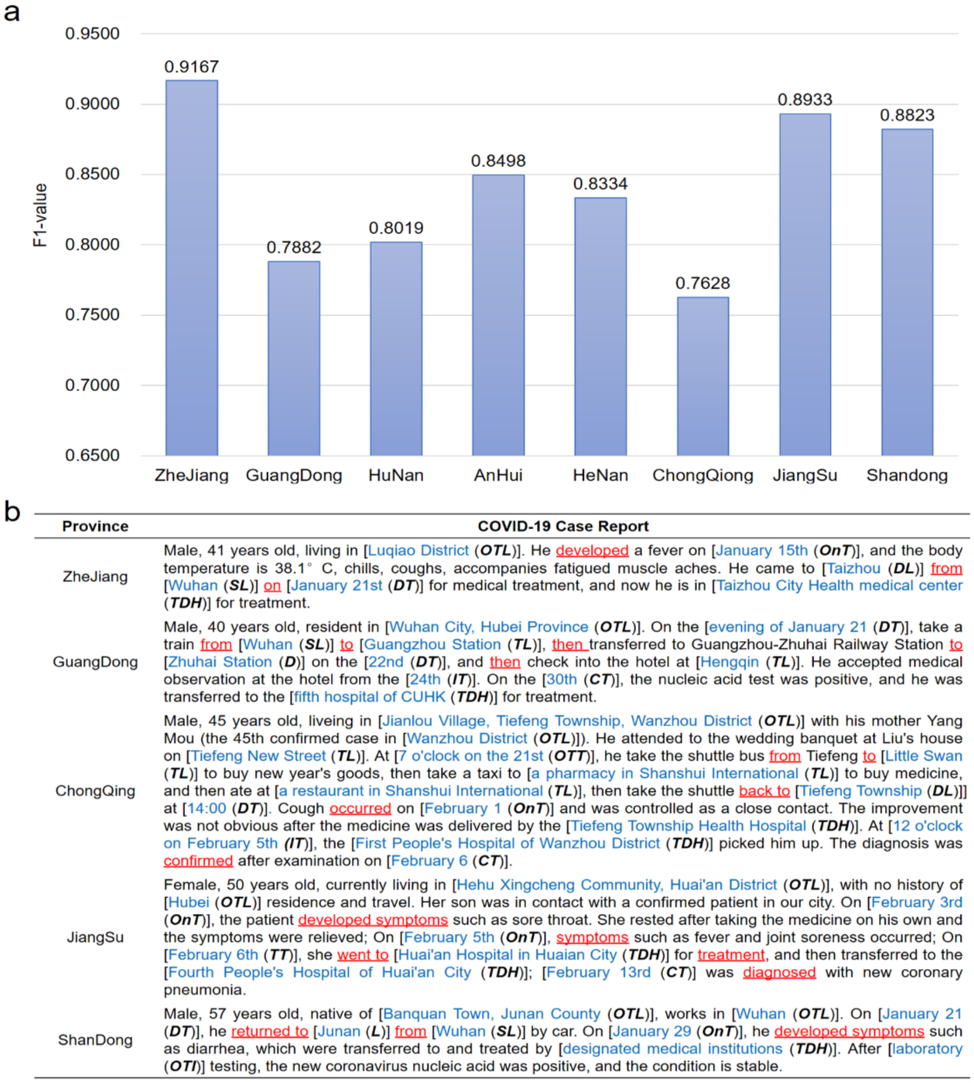


Figure 3. Performance gain across all data fields by increasing the annotation size. The CCIE is used for the named entity recognition and text classification. Blue curve in the upper panel indicates the increase in the overall F1-value as the size of annotation increases. Red curve in the bottom panel indicates the reduction in the revenue value as the size of annotation increases. The revenue value is calculated as $(after - original)/original$, with $after$ indicating the F1-value of CCIE after increasing the size of annotation data, and $original$ indicating the F1-value of CCIE before adding annotation data. When the size of annotation data is smaller than 200, the revenue value is calculated for every additional 20 annotation data; when it is larger than 200, the revenue value is calculated for every additional 100 annotation data.



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 (***OTI***)] testing, the new coronavirus nucleic acid was positive, and the condition is stable.

Figure 4. Performance of CCIE in identifying data fields from COVID-19 case reports disclosed by each province in China. (a) Eight provinces including Zhejiang, Guangdong, Hunan, Anhui, Henan, Chongqing, Jiangsu, and Shandong are selected to assess the effectiveness of CCIE in handling case reports with different natural language written styles. (b) To illustrate the data fields identified using our CCIE, the text boxes provide five examples of case reports from Zhejiang, Guangdong, Chongqing, Jiangsu, and Shandong, with the identified data fields highlighted in blue color, the trigger words highlighted in red color, and the field labels highlighted in bold black.

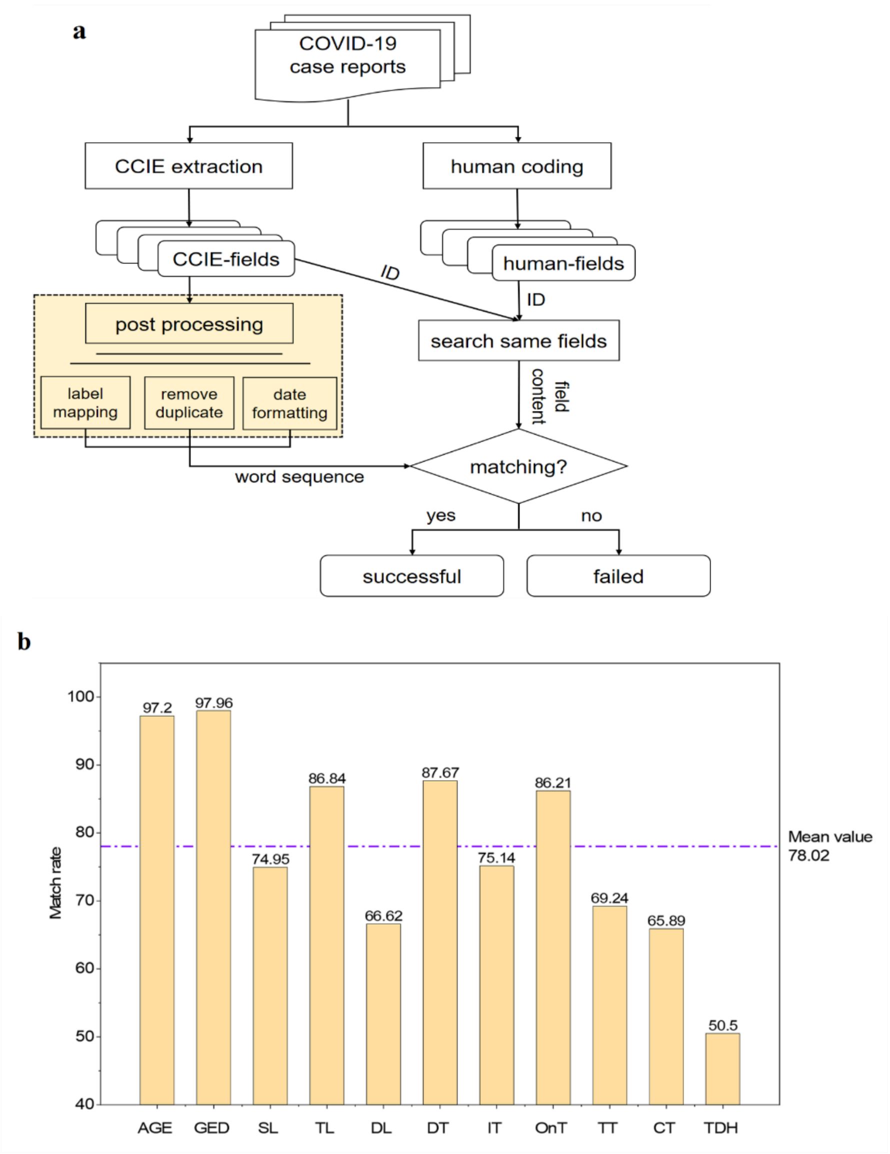
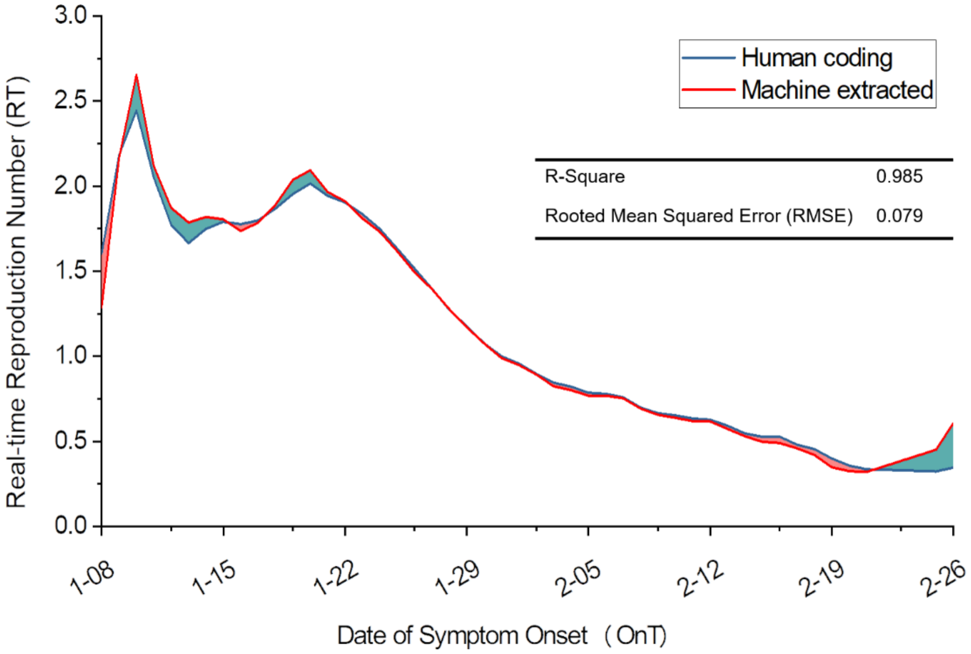


Figure 5. Cross-validation with manually extracted named entities. (**a**) The fuzzy matching method (highlighted in yellow color) is used to compute the matching rate, in which the term “label mapping” indicates the projection of named entities obtained from our CCIE into the readable fields, the term “remove duplicate” indicates the removal of duplicate values for the same data field, and the term “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 eleven data fields identified from our CCIE algorithm as compared to the gold standard results obtained by manual human coding. The mean value of accuracy averages over all data fields.



R-Square	0.985
Rooted Mean Squared Error (RMSE)	0.079

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 of manual human coding. The analysis uses COVID-19 cases with symptom onset occurring between 8 January 2020 and 26 February 2020. The R_t is estimated using ready-to-use tool [47], which is implemented in popular software including Microsoft Excel. To quantify the accuracy of CCIE in estimating R_t, the inset panel shows the R-square and the rooted mean squared error (RMSE) for the time series of R_t estimated with the two data-extraction methods.