A COVID-19 Knowledge Graph Analysis System

Dong Liu

Abstract—Nowadays, the prevention and control of COVID-19 have received significant attention. The critical point of it lies in the epidemiological surveys of patients and the further analysis of epidemiological survey reports (case reports). However, the current mainstream analysis approaches are all made manually, which is time-consuming and manpower-intensive. Therefore, a more efficient and intuitive epidemiological analysis approach is needed. I design and implement an end-to-end automated visual analysis framework for the applications for the epidemiological survey on COVID-19 named AVESA (Automated Visual Epidemiological Survey Analysis). The framework first designs and implements a deep neural network based information extraction model for case reports. Then, based on this pre-designed knowledge graph schema, the epidemiological knowledge graph is automatically constructed. Finally, I design and implement a multi-dimensional knowledge reasoning model for conducting knowledge reasoning in the complete epidemiological knowledge graph of COVID-19 and visualising the results. I conduct experiments to prove the effectiveness of this work based on the background mentioned above.

Index Terms—Epidemiological survey analysis, knowledge graph, knowledge reasoning, epidemic

1 Introduction

With the rapid development of politics and economy in human society, epidemic prevention and control has become one of the most urgent tasks for governments worldwide. It is not only related to people's health but also has a profound impact on people's livelihood and the long-term stability of countries, and it even affects the harmony and stability of international relations.

Epidemics refer to diseases produced, transmitted or infected in living organisms. Some epidemics can not only be transmitted from person to person but also from animals to humans. Also, the ways of transmission can vary widely, which makes epidemics hard to prevent and control. For example, the outbreak of severe acute respiratory syndrome (SARS) in China in 2003, hepatitis A, hepatitis B, the highly deadly viral hemorrhagic fever caused by the Ebola virus, Middle East respiratory syndrome (MERS) and the coronavirus disease 2019 (COVID-19) which is circulating all over the world [?].

The term "epidemiological survey" has gradually become familiar to the general public as China comprehensively pushed forward various measures to contain COVID-19. The epidemiological survey refers to some work done by epidemiological survey workers and scholars to control epidemics and investigate and deal with specific objects in some epidemiological fields [?]. From the perspective of epidemiology, the reason why COVID-19 can become an epidemic is that it has the same three fundamental factors as other epidemics: infection source, transmission chain and susceptible population [?]. Therefore, early detection and reporting of COVID-19, followed by timely quarantine and early treatment of confirmed patients and their close contacts, are the primary measures to contain COVID-19. Among them, early detection is the basis of other measures. Without early detection, it will be difficult to actively contain the spread of COVID-19 and carry out the dynamic zero policy. Only when the potential COVID-19 transmission factors are identified early can the source of the epidemic outbreak be stuck, so that the virus cannot spread rapidly on a large scale. The epidemiological survey is critical and meaningful to achieve early detection.

In the epidemiological survey work, the workers who collect, sort out and analyze the case reports are experts in conducting a comprehensive survey, analysis and judgment on the "sociality" of COVID-19 patients. Epidemiological survey work is similar to the consultation process before the doctor concludes a patient's condition. Epidemiological survey workers can track down people and places that need attention in COVID-19 control through various ways, such as visiting the place for investigation or inquiring about relevant people [?]. Only when the case reports of each confirmed patient are correctly arranged can experts have the opportunity to discover the transmission pattern and the potential transmission chain of COVID-19 to achieve precise control and the dynamic zero policy.

In general, the following three steps are used to record a case report of a confirmed patient: (1) Identify the source of infection of the confirmed patient; (2) Explore the transmission chain; (3) Analyze the susceptible population. Through the feedback from various channels, epidemiological survey workers can finally form a complete and detailed case report, which provides a basis for implementing COVID-19 control measures. To continue with the sorted case report, the experts have to find compelling clues about epidemic transmission from a large amount of text content, which is more challenging than collecting the original epidemiological information. Hence, the analysis of case reports is the key to epidemic control.

1.1 Contributions

The main contributions of this paper are as follows:

- I design and implement an end-to-end automated COVID-19 epidemiological analysis framework AVESA

- *Dong Liu is an undergraduate student at the School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China. E-mail: dong.liu@bit.edu.cn.*

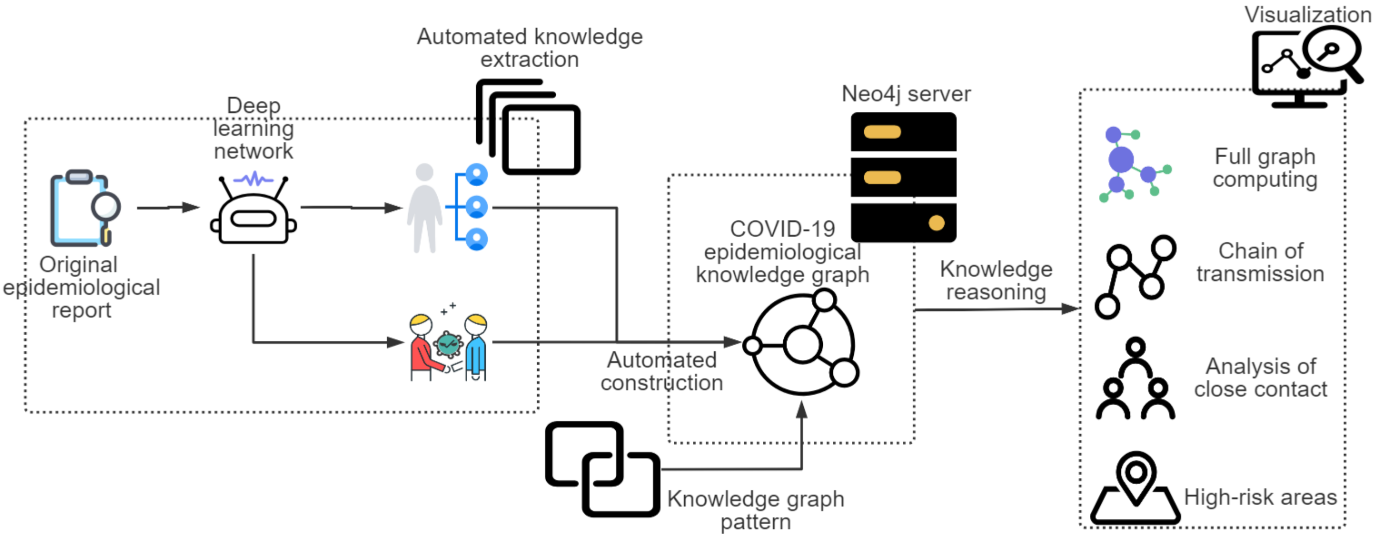

Fig. 1. Total structure of the AVESA framework.

as shown in Fig. 1, which solves the inefficiency and unintuitiveness of traditional manual epidemiological analysis.

- I design and implement a knowledge extraction model based on the combination of horizontal fusion, vertical fusion and deep learning, which solves the problem of low performance of directly transfer models trained in other fields.
- I design and implement a complete set of the automated construction process of the epidemiological knowledge graph of COVID-19 and generate an epidemiological knowledge graph with 550 nodes and 1385 edges based on existing case reports text.
- I establish fthis complete reasoning rules with different dimensions based on the epidemiological knowledge graph and present visualized results. These results provide an essential basis for epidemiological survey workers to select the next epidemic control strategies.
- I design a series of comparative experiments for applications, conduct experiments, compare results, and finally demonstrate the feasibility and advantages of the AVESA framework.

2 Related Work

The AVESA framework we propose mainly involves three fields: COVID-19 epidemiological survey analysis, natural language processing and knowledge graph. In what follows, we will introduce a brief overview of related works in these fields.

2.1 Research on COVID-19

In recent years, researches on COVID-19 epidemiological survey have shown a rapid growth trend, which are mainly divided into the following two categories:

1) Analysis of epidemiological survey results of COVID-19. Gong *et al.* [?] analyze the epidemiological survey results of a COVID-19 cluster outbreak in a public place and conclude that COVID-19 can be transmitted through contact in public places, leading to a cluster epidemic. By analyzing the epidemiological survey of a COVID-19 familial cluster outbreak, Bai *et al.* [?] conclude that familial clusters are the focus of COVID-19 control. Guo *et al.* [?] analyze the epidemiological characteristics of the COVID-19 cases and conclude that most patients who tested positive for COVID-19 had a clear epidemiological contact history. The conclusion of the above work mainly relies on the manual analysis of case reports. When the number of reports is large, this traditional manual analysis method will be time-consuming and manpower-intensive. At the same time, due to the quick spread of COVID-19, if reports could not be analyzed in time, it would be hard to obtain information on the transmission chain and close contacts, which increases the difficulty of COVID-19 control.
2) Methods of COVID-19 epidemiological survey analysis. Chen *et al.* [?] use the HanLP Chinese processing package to analyze the background information and travel records of COVID-19 patients. Wu *et al.* [?] use the Conditional Random Field (CRF) algorithm to extract information tuples from structured, semi-structured, and unstructured data published by news media and official Chinese government websites. Wang *et al.* [?] propose a network-based automated analysis and reasoning framework for COVID-19 patients' background information and travel records. They first design and implement a tuple-based neural network TMT-NN to extract epidemiological entities and relations from case reports. The above work is all post-event analysis of the COVID-19 epidemiological survey, and there are challenges in automated text analysis of COVID-19 case reports (such as when close contacts are unknown). In addition, there is still much room for improvement in the extraction performance of epidemiological information, potentially leading to a more effective COVID-19 control decision.

2.2 Natural Language Processing

With the introduction of Google's [?] Transformer, the slow training of recurrent neural networks has been improved. Moreover, the self-attention mechanism can then be used to achieve fast parallel computing. The advent of BERT [?] based on Transformer has promoted the effectiveness of solving various vital tasks in the field of natural language processing to a new level. For instance, Jiang *et al.* [?] propose a BERT-BiLSTM-CRF based causality extraction model. Zeng *et al.* [?] propose a BERT and joint learning based model for named entity recognition (NER) in referee documents. Liu *et al.* [?] propose a hybrid method based on BERT and BiLSTM for unbalanced text sentiment analysis of network public opinion. Yan *et al.* [?] propose an online alarm recognition method for power grid regulation based on BERT-DSA-CNN and knowledge base.

Thanks to the proposal of BERT, natural language processing technology has also been widely used in the medical field. Xu *et al.* [?] propose a medical-aided diagnosis model based on BiLSTM and BERT, the core principle of which is to incorporate global information into the extraction of local features to obtain more local features in the text. Xu *et al.* [?] propose a model combining the pretrained language model BERT and BiLSTM for biomedical named entity recognition. Pan *et al.* [?] propose a BERT based model that can comprehensively consider medical text features and SQL query structure to automatically convert medical text into SQL queries for electronic medical records. Zhou *et al.* [?] propose a BERT based model for medical question answering.

Researchers have made full use of natural language processing technology to achieve various tasks related to COVID-19. Khadhraoui *et al.* [?] propose a pre-trained language model CovBERT based on BERT to automate the task of classifying research documents. Chintalapudi *et al.* [?] use a deep learning model to perform sentiment analysis on online public opinion. Dong *et al.* [?] propose a BERT based intelligent question answering system for COVID-19. Jiang *et al.* [?] propose a BERT based method to automatically extract entity knowledge of major public health infectious disease events. The above works have laid a good foundation for the work of this paper.

2.3 Research on Knowledge Graph

The core idea of a knowledge graph is to express knowledge and data explicitly with a graph. Its research covers a series of well-known interdisciplinary technologies, such as knowledge representation and reasoning, information retrieval and extraction, data mining and machine learning [?].

The applications of knowledge graphs include large-scale data integration, data management and insights extraction from various data sources [?]. In 2012 Google released the first knowledge graph in the modern sense for search-related business [?]. In the application of knowledge graphs, using the knowledge abstraction based on the graph model has the following two essential advantages. First, the graph model provides a concise and intuitive abstraction, and the relationship between nodes is reflected by edges, which is very consistent with the close contact relationship in the case report. Second, the graph model does not require developers to determine the data storage and representation method at the early stage of design. Especially in the context of having incomplete knowledge, it is highly scalable.

Knowledge graph technology has been widely used in the medical field. Hu *et al.* [?] develop a web application DGLinker for predicting disease-gene associations, allowing users to utilize biomedical information from various biological databases, generate knowledge graphs and use machine learning to predict new disease-related genes. Chai *et al.* [?] construct a medical knowledge graph of thyroid diseases and apply it to intelligent medical diagnosis. Ansong *et al.* [?] explore how to use knowledge graph technology to help solve the problem of insufficient disease diagnosis data and interpretability to build reliable and effective disease diagnosis functions in medical training systems. Zhang *et al.* [?] use graph convolutional neural network modelling to assist radiology image report generation and combined knowledge graph technology to achieve feature learning for modelling the relationship between diseases.

Researchers have used knowledge graph technology to achieve various tasks related to COVID-19. Si *et al.* [?] retrieve journal literature related to the prevention and control of COVID-19 from CNKI and use the knowledge graph as a data source to perform cluster analysis from multiple perspectives such as keywords, authors, and cooperative institutions, and draw a visual map. Yang *et al.* [?] construct a knowledge graph of COVID-19 scientific literature, a knowledge graph of traditional Chinese medicine treatment, and a knowledge graph of Western medicine treatment and realize the integration of knowledge graphs. Ren *et al.* [?] have initially implemented an intelligent question-answering system based on the knowledge graph of COVID-19, which can provide real-time consultation services for COVID-19 disease. The above work has laid a good foundation for the work of this paper.

3 Problem Definition and Proposed Method

In this section, I introduce some background knowledge about the case report and define the AVESA framework.

3.1 Problem Definition

The main work of this paper is to use the unstructured and semi-structured COVID-19 case report text as input and provide an automated visualized analysis method based on deep neural networks and other artificial intelligence technologies.

Specifically, a standardized case report text for COVID-19 should contain the following parts: (1) Background. (2) Symptoms and preliminary diagnosis. (3) Investigation of source contagion. (4) PCR test. (5) Results of investigation. (6) Primary close contacts and secondary close contacts. (7) Measures taken.

Among them, Part 1 is the basic information about the patient. This part mainly records the basic personal information of the patient (e.g. name, gender, age, place of work, place of residence) and their direct or close relatives. Since the basic information of patients can preliminary help determine the potential transmission places of COVID-19 and the information about relatives can preliminary help

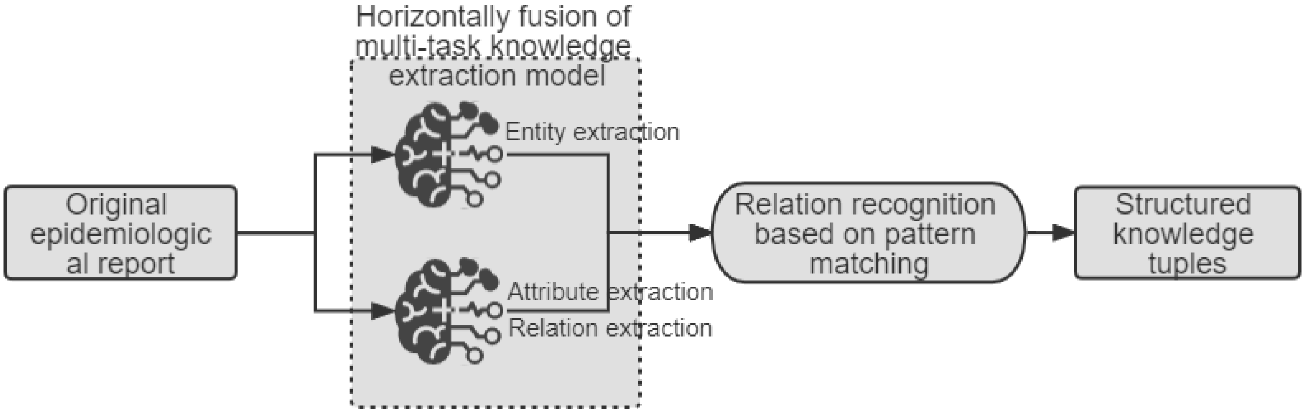
Horizontally fusion of multi-task knowledge extraction model
Original epidemiologic al report
Entity extraction
Attribute extraction
Relation extraction
Relation recognition based on pattern matching
Structured knowledge tuples

Fig. 2. Flow diagram of knowledge extraction of case reports.

find out close contacts with a higher risk of infection, so having an accurately analyzed Part 1 is very crucial; Part 2 mainly records the patient's onset time and health situation; Part 3 details the patient's activity records and contacts in the 14 days before confirmed, which is significant to the epidemiological work. It is also an important starting point and basis for workers to conduct epidemiological analysis; Part 4 mainly includes the time and results of polymerase chain reaction tests; Part 5 is the conclusion of the preliminary investigation; Part 6 is the determination of primary close contacts and secondary close contacts; Part 7 describes the measures taken, including the quarantine of patients and their close contacts and the disinfection of the locations visited by patients; Part 8 is the suggestions for the subsequent stage work. Except for Part 1 and Part 3, the rests are summaries or brief records of the results of the preliminary epidemiological survey made by epidemiological survey workers on case reports using traditional manual methods. Therefore, this paper focuses on the contents of Part 1 and Part 3. The accuracy and effectiveness of this work are verified by taking the conclusions of the traditional COVID-19 epidemiological survey analysis as a reference.

Although the automated visualization of COVID-19 epidemiological analysis can provide more effective help for the current epidemic control, there are still problems in the current work on this issue:

1) Accurate semantic understanding of unstructured and semi-structured case reports is challenging. There are many types of entities that need to be recognized in the case report. In addition to person and place names, it is also necessary to accurately recognize the time of the event and the relation between entities (e.g. lineage). Due to the confidentiality of COVID-19 epidemiological analysis, academia and industry do not yet have complete labelled datasets. Therefore, it is impossible to use existing open source datasets to train models to achieve accurate and efficient analysis.
2) It is challenging to construct a complex epidemiological knowledge graph of COVID-19 and reason adequate information on the transmission chain, potentially infected persons and dangerous areas. Due to the long incubation period and strong contagion of COVID-19, the case report records a long period of personal activities involving people and places, which is very complicated. Therefore, designing a suitable knowledge graph pattern and reasonable knowledge reasoning rules is highly important for rationally constructing a knowledge graph and quickly obtaining adequate information from the vast and complex epidemiological knowledge graph.

3.2 Proposed Framework

Fig. 1 shows the entire structure of the AVESA framework. As shown in the figure, the work of this paper is mainly divided into three stages:

1) Construction of the knowledge extraction model of case reports.
2) Construction of the knowledge graph.
3) Construction of the knowledge reasoning model and the visualization of the results.

These three stages are described in detail respectively in Section 3.3, Section 3.4 and Section 3.5.

3.3 Automated Analysis Model

As shown in Fig. 2, knowledge extraction and relation recognition are two main works in this stage.

After a comprehensive and detailed analysis of 21 case reports (the original texts obtained directly from the Chinese Center for Disease Control and Prevention) and the characteristics of the content of Part 3, I design and implement a suitable deep neural network model to analyze the case reports automatically.

Next, I will introduce the three sub-modules of the automated analysis model of the case reports.

3.3.1 Vertical Fusion Entity Extraction

Researchers have demonstrated the superiority of using pre-trained language models as encoders for tasks related to natural language processing [**?**].

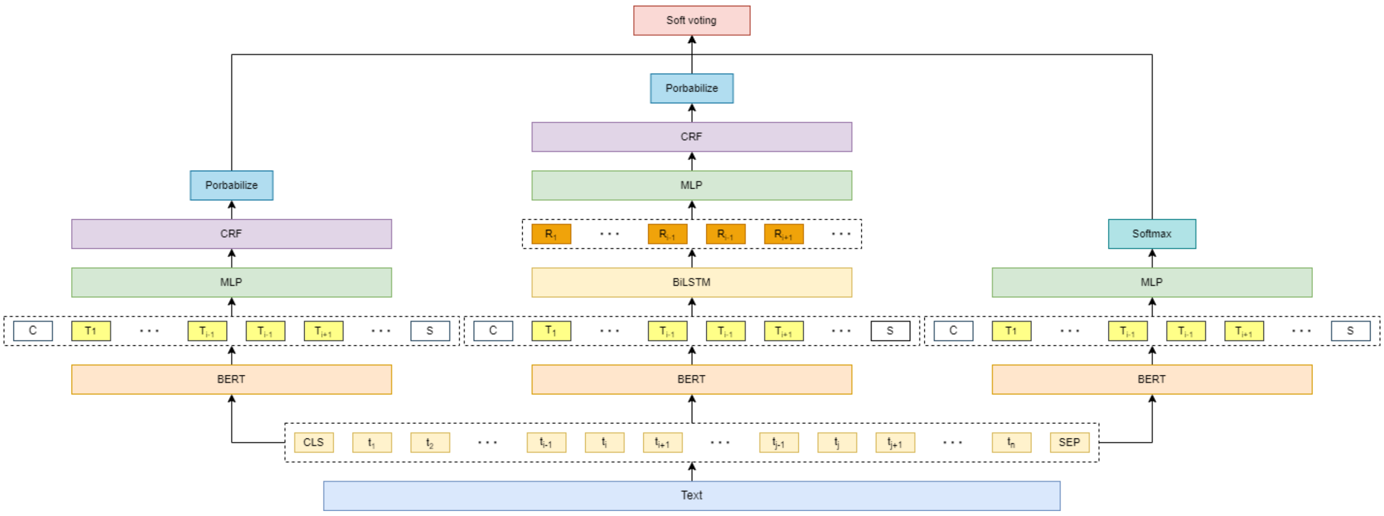
Soft voting
Porbabilize
CRF
MLP
R_1 ··· R_{i-1} R_i R_{i+1} ···
BiLSTM
Porbabilize
CRF
MLP
Softmax
MLP
C T1 ··· T_{i-1} T_i T_{i+1} ··· S
BERT
BERT
BERT
CLS t_1 t_2 ··· t_{i-1} t_i t_{i+1} ··· t_{j-1} t_j t_{j+1} ··· t_n SEP
Text

Fig. 3. Ensemble model for epidemiological knowledge extraction.

The Long Short-term Memory (LSTM) model has a strong ability to extract long sequence features, and the bidirectional LSTM (BiLSTM) can extract bidirectional text information, further improving the model's feature extraction ability.

The Conditional Random Fields (CRF) model helps to calculate the global optimum of the entire sequence, which can obtain the predicted sequence with the highest probability by decoding the encoded sequence. The combination of bidirectional LSTM and CRF, the BiLSTM-CRF model, is effective on named entity extraction (NER) tasks [**?**].

Ensemble learning improves the prediction performance of a single model by training multiple models and combining their predictions.

The most common method used in ensemble learning, soft voting, counts the average of the prediction probabilities of multiple models in a certain category and selects the category with the highest average as the final prediction result.

In this work, I use BERT-MLP, BERT-CRF and BERT-BiLSTM-CRF as the three basic models and soft voting to complete the NER task. Fig. 3 shows the process of using BIO labelling rules to extract epidemiological entities. Specifically, it is divided into three stages: Encoding, character category probability calculation, model ensemble and labelling. The specific content is described below.

Suppose a text sequence has n characters, these n characters will be represented as $T = \{t_1, t_2, ..., t_n\}$ as input to the model.

Encoding In all three models, I use the BERT model pre-trained on the Chinese corpus to encode the text sequence. Specifically, given the input sequence $T = \{t_1, t_2, ..., t_n\}$, the special identification characters [CLS] and [SEP] are inserted into the beginning and the end of the sequence respectively, and the BERT-encoded output $T^B = \{T_1, T_2, ..., T_n\}$ is regarded as the representation of the sequence. The BERT-BiLSTM-CRF model also uses BiLSTM to further encode the sequence. The obtained output $R = \{R_1, R_2, ..., R_n\}$ can then also be regarded as the representation of the sequence T.

Character category probability calculation In this stage, in the BERT-MLP model, the multi-layer perception (MLP) and the softmax layer map the sequence representation to the label sequence. If there are m categories, then the vector $\hat{y}_i = \{\hat{p}_1, \hat{p}_2, ..., \hat{p}_m\}, (\sum_{i=1}^{m} \hat{p}_i = 1)$ represents the probabilities of this character on each category. In the other two models, due to the determinacy of CRF decoding, it cannot naturally participate in soft voting. Therefore, I heuristically probabilize the sequence of labels obtained by decoding the CRF. For each character, I give the decoded category on that character a probability of α, and 0 for the other classes. This probabilistic representation meets the requirements of soft voting so that the model decoded by CRF can still participate in soft voting with other models.

Ensemble and labelling According to soft voting, the label probability matrix obtained by the three models is summed and averaged in this stage, and the vector $\hat{y}_i^V = \{\hat{p}_1^V, \hat{p}_2^V, ..., \hat{p}_m^V\}$ in the final sequence $\hat{y}^V = \{\hat{y}_1^V, \hat{y}_2^V, ..., \hat{y}_m^V\}$ is the probability vector of the category of each character.

These three models do not share parameters in the task and perform model training independently.

3.3.2 Horizontal Fusion Multi-task Knowledge Extraction

Knowledge extraction tasks can be divided into three categories: entity extraction, attribute extraction and relation extraction. According to the data characteristics of the case reports, I design a vertical fused deep learning ensemble model described in detail in Section 3.3.1 and use different datasets to train on these three tasks.

The first model mainly focuses on entity extraction, while the second multi-task knowledge extraction model mainly focuses on attribute extraction and relation extraction. After the automated analysis of the same case report, the two models will adopt the sequential union method to achieve horizontal fusion. For example, for the following text *Jenny, female, age 37, address: Zhongguancun. Husband Watson, male, farmer*, the prediction result of the first model is *[(Jenny, name), (Zhongguancun, residence), (Watson, name)]*, the prediction result of the second model is *[(female, gender), (37, age), (Husband, social relation), (male, gender)]*.

After horizontal fusion, the output of the model is *[(Jenny, name), (female, gender), (37, age), (Zhongguancun, residence), (Husband, social relation), (Watson, name), (male, gender)]*.

The multi-task knowledge extraction model of horizontal fusion is used to extract entities, attributes and relations, which lays the foundation for subsequent relation recognition.

3.3.3 Pattern Matching Based Relation Recognition

After extracting entities, attributes and relations from the original text through the deep learning model and ensemble learning methods, I need to utilize the model's output in Section 3.3.2 above to obtain the essential information for building a knowledge graph. This way, the following two types of relations are identified and constructed. For example, for the results of the multi-task knowledge extraction model: *[(Jenny, name), (female, gender), (37, age), (Zhongguancun, residence), (Husband, social relation), (Watson, name), (male, gender)]*, there are two types of problems:

1) Whether there is a relation between different entities, and if so, what kind of relation? The problem is identifying the relation between ’A’ and ’D’ (that is, ’D’ is ’A”s husband).
2) Which entity or relation should the extracted attribute belong to? The problem is identifying that gender ’female’ and age ’xx’ are attributes of ’A’, while gender ’male’ is an attribute of ’D’.

In order to solve the above two problems, I investigate the three most popular relation recognition and construction models:

1) Pattern matching model: This model constructs general pattern matching rules by analyzing the input corpus and adding domain expert knowledge. Using these pre-defined templates, relations can be identified and constructed in a new corpus.
2) Statistical learning model: This model uses classical machine learning algorithms (e.g. support vector machines, Bayesian algorithms) to convert the problem of relation recognition into a classification problem and then build relations in a new corpus.
3) Neural network model: The model uses different neural network architectures for training to learn the semantic relations knowledge in a corpus to predict possible relations in a new corpus.

Although the latter two models have been successfully applied in some natural language processing fields, they require sufficiently complete annotated datasets as data support for their training. However, because of the lack of mature and complete annotated datasets in the field of COVID-19 case report text analysis and the fact that the case report text follows the format requirements of a unified standard, employing the pattern matching model can maximize the prior knowledge of experts and achieve a relatively high level of accuracy. Therefore, I use the pattern matching model with the following rules:

1) For the identification and construction of ’lineage’ relation: If a ’lineage’ entity is extracted from a certain sentence of the case report text, it is considered an attribute of the new ’lineage’ relation. The infected person of the report belongs to one end of the relation, and the other end should be the person corresponding to the ’name’ entity closest to the ’lineage’ entity extracted from the current sentence.
2) For the identification and construction of ’locate’ relation: If a ’time’ entity is extracted from a sentence in Part 3 of the case report text, it is considered an attribute of the new ’locate’ relation. One end of it should be The location extracted from the current sentence, and the other end should be the person corresponding to all the ’name’ entities extracted from the current sentence and the infected person of the report.
3) For the identification and construction of ’attribute’ relation: If an attribute entity (e.g. age, gender) is extracted from a sentence in Part 1 of the case report text, it is considered to be the attribute corresponding to the subject of the sentence.

After the automated analysis and knowledge extraction of the COVID-19 case report text, the extracted knowledge should be standardized so that the terms are consistent in the entire knowledge graph. Then I can use the standardized terms as input for constructing the knowledge graph.

3.4 Construction of Knowledge Graph

There are two main components in a knowledge graph, nodes and relations. A node is a basic data unit, and each node has a unique ID; A relation is another component that indicates whether there is a connection between two nodes. More specifically, in order to construct a complete epidemiological knowledge graph of COVID-19, I should focus on the following two parts:

1) Pattern layer: The pattern layer represents a knowledge graph's storage structure and organizational form. It commonly contains the definition of each node and relation in a knowledge graph, including the information of types and attributes.
2) Data layer: The data layer refers to the specific data stored and organized by each node and relation in a knowledge graph based on the pattern layer.

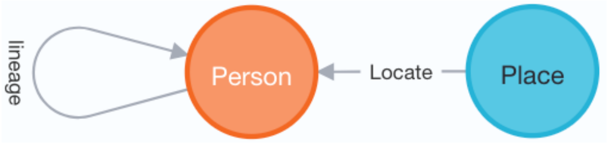

Fig. 4. The pattern layer of the epidemiological knowledge graph.

3.4.1 Design of Pattern Layer

Before constructing the epidemiological knowledge graph of COVID-19, I first design the pattern layer as shown in Fig. 4.

A complete epidemiological knowledge graph includes at least two types of nodes: *Person* and *Place*. Each node labelled ’person’ should include attributes such as name, gender, age, and residence. It should be noted that not all case report texts indicate all the attributes of the relevant personnel, so the other attribute values may be empty except for the name. Each node labelled ’location’ should include location name, specific and other attributes.

A complete epidemiological knowledge graph includes at least two types of relations: *Lineage* and *Locate*. The relation in which each category is ’lineage’ shall include specific kinship categories, such as father, mother and husband. Each ’lineage’ relation connects two ’person’ nodes, indicating that the two people have this relationship. Each relation with the category of ’locate’ should include a specific time (e.g. 1st January). Each ’locate’ relation connects a ’person’ node and a ’place’ node, indicating that the person has been to the place at the corresponding time.

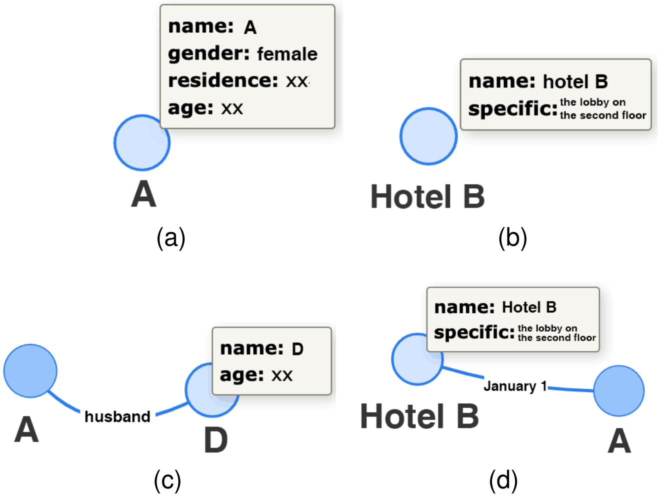

Fig. 5. Implementation of the data layer of the epidemiological knowledge graph. (a) Example of a ’person’ node. (b) Example of a ’place’ node. (c) Example of a ’lineage’ relation. (d) Example of a ’locate’ relation.

3.4.2 Design of Data Layer

When the pattern layer is constructed, the results obtained from the text analysis of the case report will be automatically mapped to the pattern to form the corresponding data layer. For example, the text of a case report is *A, female, age xx, address: xx village*, the node corresponding to the label ’person’ should be as shown in Fig. 5a.

For the text *On 1st January, A entered the lobby on the second floor of Hotel B for dinner*, the node corresponding to the label ’place’ should be as shown in Fig. 5b.

For the text *D, the husband of A, male, xx years old*, the relation with the corresponding label ’lineage’ should be as shown in Fig. 5c.

For the text *On 1st January, A entered the lobby on the second floor of Hotel B for dinner*, the relation with the corresponding label ’locate’ should be as shown in Fig. 5d.

I use Neo4j graph database [**?**] and Cypher [**?**] statement to build and store the epidemiological knowledge graph of COVID-19. I deploy the Neo4j graph database on a remote server and use the Py2neo client library to execute CREATE and MERGE statements so that all nodes and relations are inserted into the graph database. The knowledge graph information can be remotely accessed locally through Neo4j Browser.

3.5 Knowledge Reasoning Model

The complete epidemiological knowledge graph contains a large amount of complex epidemiological information. An adequate knowledge reasoning method is required to mine knowledge valuable to epidemiological survey workers from the information. Therefore, I design and implement two types of knowledge reasoning models with different dimensions.

The first type is the knowledge reasoning model based on full-graph computing. The main goal of it is to control high-risk groups and environmental communities with a greater risk of epidemic spread from a macro perspective. The second type is the transmission link model based on sub-graph matching. The main goal of it is to accurately identify the people who have the possibility of being primary close contacts and secondary close contacts and the places that need to be disinfected.

In the visualized results of the epidemiological knowledge graph, due to the detailed information in the COVID-19 case reports, the edges of the complete knowledge graph are very complicated. However, from a simple analysis, I believe that different types of relations should have different degrees of influence and importance. Based on a detailed study of the ”Novel Coronavirus Infection Pneumonia Prevention and Control Protocol (8th Edition)”, I have further communicated with experts and scholars of epidemiology. I define the initial weights for different edges according to the current epidemic control and transmission knowledge. The initial weights are divided into three grades, of which the ’lineage’ relation is the largest. If there is a ’lineage’ relation between two ’person’ nodes, the weight of the edge between them is defined as 0.8; if there is a relation between a ’person’ node and a ’place’ node, then the initial weight is 0.2; if it is finally concluded that the ’place’ node has a relation with multiple ’person’ nodes at the same time, the corresponding initial weight of the edge is defined as 0.4. In addition, in order to facilitate the subsequent establishment of a knowledge reasoning model based on full-graph computing, in the complete epidemiological knowledge graph, I have included all ’person’ nodes that are related to a ’place’ nodes at the same time, the corresponding ’place’ nodes and all ’person’ nodes who are related as the COVID-19 contact sub-graph G_c.

3.5.1 Full-graph Computing

Since COVID-19 has the characteristics of fast transmission, long incubation period, and spanning a long time and space, people who are primary or secondary close contacts of the confirmed patients have a higher risk of being infected with COVID-19. Also, the COVID-19 virus can easily remain in the places where the confirmed patients have been. It is necessary to quarantine these people and disinfect the corresponding places. Therefore, to collect the information about the people mentioned above and places accurately and efficiently, conducting full-graph computing of the epidemiological knowledge graph is necessary. Specifically, the knowledge reasoning model based on full-graph computing mainly includes the following three aspects.

Close Contacts Determination Model According to the epidemiological knowledge graph of COVID-19 and the definition and determination method of primary and secondary close contacts in the ”Novel Coronavirus Infection Pneumonia Prevention and Control Protocol (8th Edition)”, I define the concept of node similarity $S_{i,j}$ between ’person’ nodes to determine primary and secondary close contacts. The node similarity $S_{i,j}$ is mainly composed of two parts: lineage similarity $L_{i,j}$ and contact similarity $C_{i,j}$. They are calculated as follows:

$$S_{i,j} = \frac{2 * L_{i,j} * C_{i,j}}{L_{i,j} + C_{i,j}}, \tag{1}$$

the lineage similarity $L_{i,j}$ is mainly used to measure the degree of 'lineage' relation between 'person' nodes. The calculation formula is as follows:

$$L_{i,j} = \begin{cases} 0.3 & \text{(if Edge}_{i,j}\text{.type = lineage)} \\ 0 & \text{(otherwise)}, \end{cases} \tag{2}$$

the contact similarity $C_{i,j}$ is mainly used to measure the degree of contact between 'person' nodes. The calculation formula is as follows:

$$C_{i,j} = \begin{cases} 0.1 * cnt_{i,j} & \text{(if } p \text{ = true)} \\ 0 & \text{(otherwise)}, \end{cases} \tag{3}$$

where $cnt_{i,j}$ represents the number of times that i and j appear at the same place at the same time, and the condition p refers to the situation that satisfies the following formula, that is, i and j are considered to appear at the same place w at the same time,

$$\begin{aligned} p = & \text{Locate}_{i,w}.\text{name} = \text{Locate}_{j,w}.\text{name} \\ & and \\ & \text{Locate}_{i,w}.\text{time} = \text{Locate}_{j,w}.\text{time}. \end{aligned}$$

Let the initial weight of the edge between person nodes i and j be $S_{i,j}$. According to the above formulas, I can first calculate the node similarity $S_{i,j}$ between them and then construct the weighted person relation sub-graph G_p.

According to this network, I can use the strength of epidemic control in specific situations as the basis for screening weights of edges to determine a specific person's possible primary and secondary close contacts.

Risk Analysis Model I use the weighted PageRank algorithm to calculate the degree of importance of each node in the contact sub-graph G_c in the epidemiological knowledge graph. The weighted PageRank algorithm calculates the PageRank value of each node based on the number of incoming relations and the importance of the corresponding source nodes. The PageRank value of a node is proportional to the node's importance in the sub-graph. Its basic assumption is that a node's importance depends on other nodes that are linked to it. Precisely, the PageRank value of any node A can be calculated as follow:

$$PR(A) = (1-d) + d * (\frac{w_1 PR(T_1)}{C(T_1)} + ... + \frac{w_n * PR(T_n)}{C(T_n)}) \tag{4}$$

where $T_i (i = 1, 2, .., n)$ is the ith node connected to node A; $PR(T_i)$ is the PageRank value of node T_i, $C(T_i)$ is the edge number connected to node T_i (that is, the degree of node T_i), w_i is the weight of the edge between node T_i and node A, d is the damping factor, the value range is (0, 1), and according to empirical evidence, it is generally 0.85 [?]. The iterative process of calculating the PageRank value of each node of the epidemiological knowledge graph is as follows:

1) Randomly assign an initial PageRank value to each node (generally $\frac{1}{n}$).
2) Update the PageRank value of each node according to Eq. (4).
3) Repeat step 2), use this group of new PageRank values to calculate another group of new PageRank values.
4) Repeat steps 2)-3) until the PageRank value converges.

Community Detection Model I use the weighted Louvain algorithm [?] to detect the communities in the contact sub-graph G_c of the epidemiological knowledge graph. The core idea of the weighted Louvain algorithm is to maximize the modularity of the entire graph data, which measures the quality of nodes assigned to the community. A good partitioning result is when the similarity of nodes inside the community is high while the similarity of nodes outside the community is low. The formula for calculating the modularity Q is

$$Q = \frac{1}{2m} \sum_c [\sum in - \frac{(\sum tot)^2}{2m}] \tag{5}$$

where m is the total number of edges in the sub-graph, $\sum in$ represents the sum of the weights of all edges in the current community, and $\sum tot$ represents the sum of the weights of the edges connecting the community c to other communities. The goal is to maximize the modularity Q, that is, the higher the sum of the weights of the edges within the community, and the lower the weights of the edges between the communities, the better. Specifically, the iterative process of the weighted Louvain algorithm is as follows:

1) Each node is regarded as an independent community, and the initial weights of edges in the community are 0.
2) For each node, traverse all connected nodes of the node, and calculate the modularity gain brought by adding the node to the community where its connected nodes are located. Select the connected node corresponding to the greatest modularity gain to join its community.
3) Fold each community formed in step 2) into a node and calculate the weights of edges between these new nodes.
4) Repeat steps 2)-3) until the community to which each node belongs remains the same.

3.5.2 Sub-graph Matching

I design and implement a transmission chain model based on sub-graph matching, which facilitates epidemiological experts and scholars to quickly obtain primary and secondary close contacts of a confirmed patient, areas that may be potentially infected by COVID-19 virus, and the potential "company of time and space" (People whose mobile phone signals are detected near the signal of a confirmed COVID-19 case receive text instructions to report to authorities.) in the epidemiological knowledge graph. Specifically, I design fthis types of omnidirectional transmission chain rules generated based on sub-graph query matching, which basically cover the core work in the epidemiological survey, as shown in Table 1.

TABLE 1
Sub-graph Query Matching Rules

Function	Input	Output	Explanation
Screen out X's primary close contacts	X's name and a date	X's primary close contacts	X was in the same place as X's close contact after the date.
Screen out X's secondary close contacts	X's name and a date	X's secondary close contacts	X's close contact was in the same place as X's secondary close contact after the date.
Screen out the places visited by X	X's name and a date	The places visited by X	X has a 'locate' relation with the corresponding 'place' node after the date.
Screen out the people who have been to Y	The name of Y and a date	The people who have been to Y	Person who has been to Y have a 'locate' relation with the corresponding 'place' node Y after the date.

4 Experiments

In this section, I conduct three types of experiments to verify the effectiveness of each part in the AVESA framework.

4.1 Automated Analysis Experiment

4.1.1 *Datasets*

After this comprehensive research, due to the particularity of the applications, neither academia nor industry has released a complete labelled dataset for COVID-19 epidemiological analysis. Therefore, I select two different open source datasets for model training. An overview of the datasets I consider in this experiments is shown in Table 2.

CLUENER2020 [**?**] is a fine-grain NER dataset. The original data is from Sina News RSS. For this application, I screen out fthis types of label categories for training: name, position, company, and government. Except for name entities, the last three label categories belong to place entities. This dataset is used in the entity extraction sub-task.

ECR-COVID-19 [**?**] is a dataset of epidemiological case reports with entity labelling which can be used for information extraction. For this application, I screen out eight types of label categories for training: age, gender, residence place, date, end date, location, spot and social relation. Among them, age, gender and residence place are the attributes of people; start time and end time correspond to the time attribute of 'locate' relation; social relation corresponds to the attribute of 'lineage' relation. This dataset is used in the multi-task extraction sub-task.

4.1.2 *Experimental Settings*

For two different sub-task extraction models, I train the same epochs in the same environment and then compare the precision, recall, and F1 scores.

For entity extraction sub-task and multi-task extraction sub-task, the models I train are both: BERT-MLP, BERT-BiLSTM-CRF, BERT-CRF and the ensemble model. The hyper-parameters, experimental environment, hardware configuration parameters and datasets used in training are consistent.

4.1.3 *Experimental Results*

Table 3 reports the results of AVESA on the entity extraction sub-task. It can be seen that the F1 score of the vertical fusion model reaches 85.12%, and the recall score reaches 87.42%, which are further improved than other individual models. The precision score is just a litter higher than the BERT model, which reaches 82.94%.

Table 4 reports the results of AVESA on the multi-task extraction sub-task. The model achieved the best scores on all three metrics on the test set.

I infer that the reason why the precision score of AVESA is not much higher than the BERT model is that the ensemble process may break the consistency of a span. More explicitly, strong false-positive predictions can badly influence the precision score.

4.2 Full-graph Computing Experiment

4.2.1 *Close Contacts Determination Experiment*

Based on the constructed epidemiological knowledge graph of COVID-19, the first step is to compute the similarity between 'person' nodes with the proposed node similarity computing method to obtain the person relation sub-graph G_p and visualize the results. The threshold is set to divide the level of close contacts and locate the relevant groups.

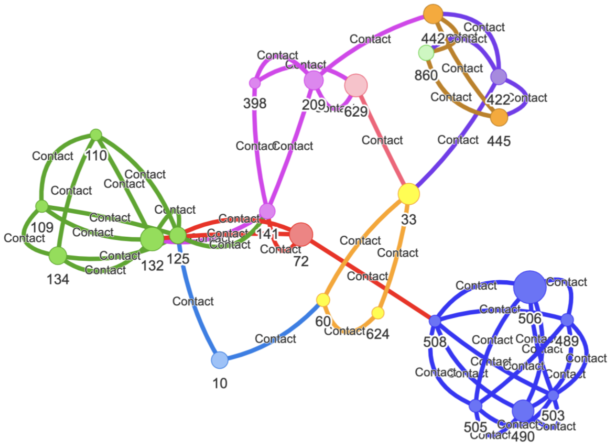

Fig. 6. COVID-19 transmission graph.

I extract the contact sub-graph from the complete epidemiological knowledge graph, screen out the patients confirmed with COVID-19 from the contact sub-graph, and obtain the transmission graph as shown in Fig. 6. It can be seen that the network integrates the knowledge belonging to different case reports. After manual verification, it is found that its transmission graph is accurate and effective.

TABLE 2
Statistics of the Datasets Used

Dataset	Rel.	Sent.	Ent.	Sent.	Ent.	Sent.	Ent.
		Train		Validation		Test	
ECR-COVID-19†	25	1,811	28,085	226	3,548	227	3,577
ECR-COVID-19	8	1,811	15,238	226	1,899	227	1,899
CLUENER2020†	10	10,748	23,338	1,343	2,982	-	-
CLUENER2020	4	10,748	8,987	1,343	1,168	-	-

† Original datasets.

TABLE 3
Entity Extraction Sub-task Comparison Results(%)

Model	Precision	Recall	F1
		Validation	
BERT	82.74	83.76	83.24
BERT-CRF	82.01	85.37	83.65
BERT-BiLSTM-CRF	81.78	86.04	83.86
AVESA	**82.94**	**87.42**	**85.12**

TABLE 4
Multi-task Extraction Sub-task Comparison Results(%)

Model	Precision	Recall	F1
		Test	
BERT	91.08	92.35	91.71
BERT-CRF	90.43	93.13	91.76
BERT-BiLSTM-CRF	90.36	92.89	91.61
AVESA	**91.11**	**93.51**	**92.29**

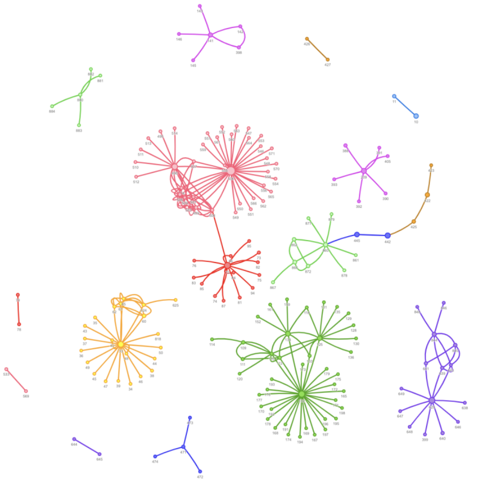

Fig. 7. The sub-graph of the person relation sub-graph with a threshold of 0.3.

I get the sub-graph when the threshold of G_p is set to 0.3, as shown in Fig. 7. After manual verification, the close contacts in the graph basically conform to the conclusion drawn by manual analysis.

I also screen out a representative transmission sub-graph for analysis, as shown in Fig. 8: According to the original case report of patient 860, patient 860 was quarantined as a close contact of patient 445 and was confirmed with COVID-19 during the quarantine period; patient 445 was the wife of patient 442, and the transmission was caused by the familial cluster outbreak; patient 425 is the neighbthis of the patient 442 and is determined as a close contact due to their direct face-to-face contact; patient 422 is the child of patient 425. To sum up, the results obtained through knowledge reasoning are consistent with the manual analysis results of case reports.

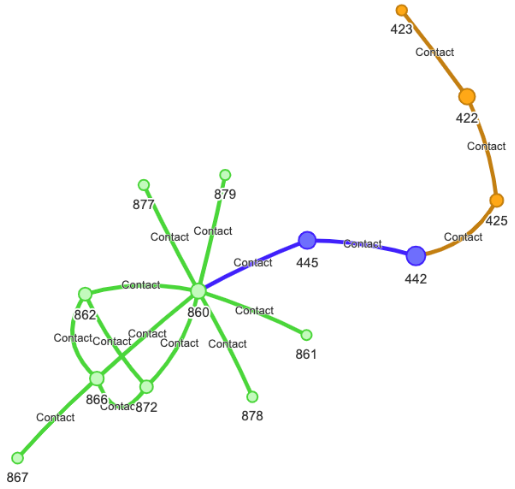

Fig. 8. A case study of the transmission chain.

4.2.2 Risk Analysis Experiment

Based on the constructed epidemiological knowledge graph of COVID-19, the first step is to construct a contact sub-graph. Based on the sub-graph, I use the control variable method to verify whether the edge weights in the contact sub-graph have an impact on the results of the PageRank algorithm. After verification, the results of the algorithm with better performance will be further analyzed in combination with the actual results.

I use the Top_Recall indicator to measure the effectiveness of the PageRank algorithm. I use the algorithm to compute the PageRank value of all ’person’ nodes and then sort them by decreasing the PageRank value. The value of Top_Recall_n% represents the rank of the last ’person’ node when the algorithm recalls n% of the confirmed patients. The lower the value of Top_Recall_n%, the more capable the algorithm is in reasoning the population with a high risk of infection. In this experiment, I set n as 25, 50, 75 and 100, respectively.

Table 5 reports the results of two algorithms. As can be seen, when n is 25, that is, at least 25% of the confirmed patients need to be recalled, for the weighted PageRank algorithm, at least the top 9 ’person’ nodes are needed;

TABLE 5
Experimental Results of PageRank Algorithm

Top_Recall	Weighted PageRank Algorithm	PageRank Algorithm	delta
25%	9	11	18.18%
50%	17	18	5.56%
75%	30	33	9.09%
100%	44	50	12.00%

for the PageRank algorithm, at least the top 11 'person' nodes are needed. The fthis groups of results show that the PageRank algorithm performs well in distinguishing the importance of 'person' nodes. Moreover, the effect is further improved after introducing the weights based on expert experience. Compared with the unweighted results, the average value of Top_Recall decreases by 11.21%.

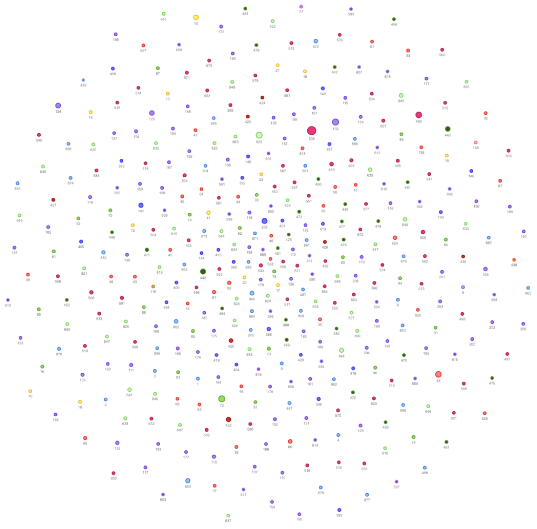

Fig. 9. Person node graph.

I conduct further analysis of the results of the weighted PageRank algorithm. Fig. 9 shows all the 'person' nodes. The node's size is determined by the PageRank value, which indicates the importance of the corresponding person in the COVID-19 epidemiological survey. After comparing with the manual analysis results, it can be concluded that people corresponding to the top ten 'person' nodes are all confirmed with COVID-19, and the degree of contact with COVID-19 patients is also highly consistent with the conclusion drawn in the node similarity experiment in Section 4.2.1. For example, according to the results of expert results, 128 primary close contacts and 50 secondary close contacts of patient 506 with the highest PageRank value are found; 123 primary close contacts and 50 secondary close contacts of patient 132 with the second highest PageRank value are found; 93 primary close contacts and 85 secondary close contacts of patient 72 with the third highest PageRank value are found;

Fig. 10 shows all the 'place' nodes. The node's size is determined by the PageRank value, which indicates the importance of the corresponding place in the COVID-19 epidemiological survey. After comparing the manual analysis results, it can be concluded that the top 10 places are all critical for disinfection or in the list of medium or high-risk areas in the COVID-19 control work. For example, place 122, with the highest PageRank value, was visited by dozens of confirmed patients.

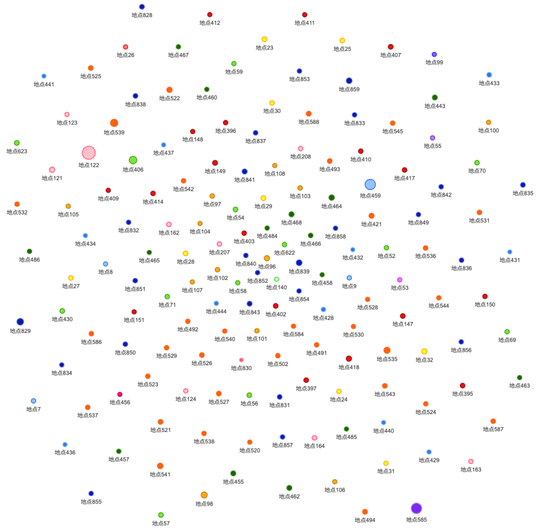

Fig. 10. Place node graph.

All the results shown above are sufficient to prove the effectiveness of the weighted PageRank algorithm.

4.2.3 *Community Detection Experiment*

Based on the constructed epidemiological knowledge graph of COVID-19, the first step is to construct a contact sub-graph. Based on the sub-graph, I use the control variable method to verify whether the edge weights in the contact sub-graph impact the Louvain algorithm's results. I also conduct a comparison experiment with the classical label propagation algorithm [?] to verify the advantage of the Louvain algorithm in this work. After verification, I further analyze the results of the algorithm with good performance in combination with the manual analysis results of case reports.

I use the community difference rate to measure the effectiveness of the community detection algorithm. The community difference rate refers to the proportion of the difference between the number of communities detected by the algorithm and the actual number of communities obtained by manual analysis. The smaller the value is, the more influential the algorithm is.

TABLE 6
Experimental Results of Community Difference Rate

Label Traversing Algorithm	Weighted Label Traversing Algorithm	Louvain Algorithm	Weighted Louvain Algorithm
+132%	+123%	+9%	+4.34%

Table 6 reports the results of fthis algorithms. As can be seen, the introduction of weights has a certain improvement on both the label propagation algorithm and the Louvain algorithm. Louvain algorithm takes edge weights based modularity as the optimization basis, which is suitable for the application of community detection in this work. And experimental results show that the Louvain algorithm performs much better than the label propagation algorithm.

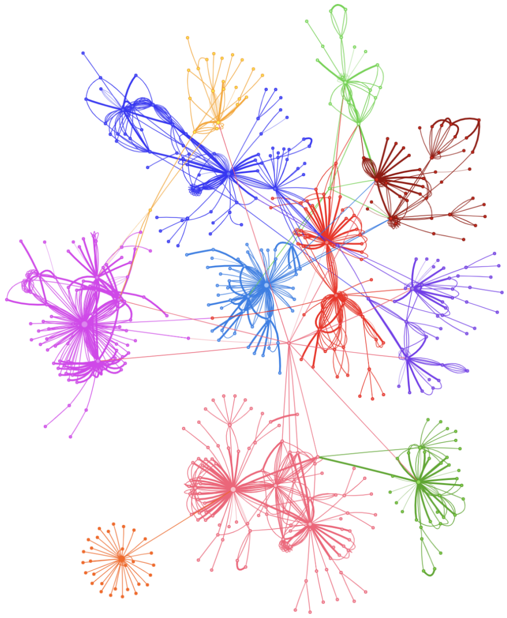

Fig. 11. Community graph.

I further analyze the results. As shown in Fig. 11, different communities divided by the Louvain algorithm are distinguished by colour. The result of community division is basically consistent with the actual contact situation. The nodes of the same colthis are basically people or places with direct relation, which appear in the same case report.

4.3 Sub-graph Matching Experiment

I experiment with the fthis rules respectively, and the results are as follows:

1) For patient 506, the primary close contact results after 3rd January are shown in Fig. 12a. After comparison, the results are entirely consistent with manual analysis, and the average search time is just about 500 milliseconds, which greatly improves the efficiency of epidemiological survey workers in finding close contacts of a specific patient.
2) For patient 445, the secondary close contact results after 5th January are shown in Fig. 12b. After comparison, the results are basically consistent with manual analysis, and the average search time is just about 1 second.
3) For patient 490, the places visited after 31st December are shown in Fig. 12c. After comparison, the results are entirely consistent with that of manual analysis. According to the information shown in the figure, I can more intuitively see the frequency of contact between patient 490 and these places.
4) For place 421, the people who visited it after 7th January are shown in Fig. 12d. After comparison, the results are basically consistent with manual analysis, and the average search time is just about 300 milliseconds.

5 Discussion

This section mainly discusses the experimental results systematically and then analyzes the advantages and areas for improvement of this work.

5.1 Advantages

In the automated analysis model proposed in this paper, the vertical fusion model performs best on all metrics I use. I can further vote and correct the output results of different models, which significantly enhances the robustness of model output. Meanwhile, the horizontal fusion multi-task knowledge extraction model divides different sub-tasks into different models for orientation training to perform their duties and improve the ability of knowledge extraction.

The pattern layer definition and automated construction process of the knowledge graph proposed in this paper are suitable for the epidemiological survey, laying a good foundation for accurate knowledge reasoning.

The multi-dimensional knowledge reasoning model proposed in this paper includes the reasoning rules needed by epidemiological survey workers. The manual analysis results also corroborate the effectiveness of the model. And sometimes, the reasoning model can find out some potential epidemic transmission risks ignored by human analysis.

5.2 Limitations

The work in this paper has the following two limitations worthy of further improvement:

1) Restriction of preset patterns in the relation recognition model: Identifying the relation between entities and the corresponding relations between attributes and entities in this work relies on fixed pattern matching rather than a more flexible deep learning model, mainly due to the lack of Chinese datasets applicable to this work. To alleviate this defect, the pattern selection in this paper adopts the guidance of expert knowledge. Therefore, if the writing of case reports follows the unified specification, the preset pattern can cover most of the needed relations in the case reports.
2) Semantic understanding limitations in the knowledge graph construction process: the following scenarios may occur in the original case report text. (a) Mistyping, a typical case is to mistype the names of people, which may be considered as belonging to different people in the process of automated knowledge graph construction; (b) Missing subject, a typical case is to omit the names of confirmed patients or other people when recording the epidemiological information, and only describe the related events, and it is more difficult to make accurate inferences based on the contextual semantic information; (c) Entity ambiguity, a typical case is to describe the same entity with different words, such as sometimes using the exact address and sometimes using the abbreviation for the same location, which may be considered as belonging to different locations in the process of automating the construction of knowledge graph.

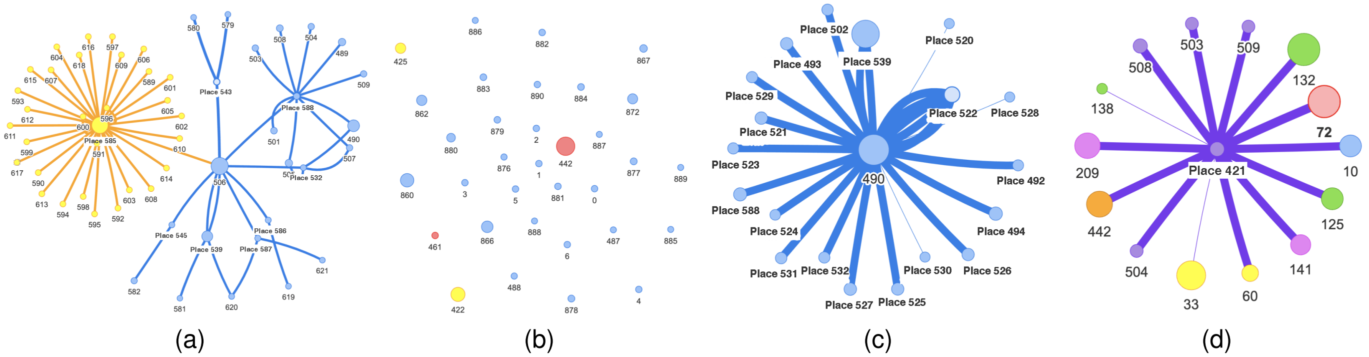

Fig. 12. (a) Partial primary close contact graph of patient 506. (b) Partial secondary close contact graph of patient 445. (c) Partial visited place graph of patient 490. (d) Partial contact graph of place 421.

6 Conclusion

In this paper, I study the problem of automated analysis of case reports. To address the challenging problem, I propose the automated visual epidemiological survey analysis (AVESA) framework. In order to accurately extract various epidemiological entities in the case report, I design a vertical fusion entity extraction model and a horizontal fusion multi-task knowledge extraction model. Next, to extract the relation between entities and the relation between attributes and entities, I design a pattern matching based relation recognition model. Moreover, I solve the problem of constructing an epidemiological knowledge graph of COVID-19. According to the expert knowledge of COVID-19, I construct a knowledge graph pattern layer with ’person’ and ’place’ as the core entities and describe the relation between entities with ’lineage’ and ’located’ relations. I use the Neo4j graph database to store specific knowledge graph data. Based on the contents of 21 original case reports provided by the Chinese Center for Disease Control and Prevention, I construct an epidemiological knowledge graph including 550 nodes and 1385 relations, which converts unstructured text content into a clear and intuitive graph model. Finally, I design a full-graph computing based knowledge reasoning model and a sub-graph matching based transmission chain model. I use the Neovis library to realize the final visualized display of the case reports.

I conduct experiments on each part of the AVESA framework and discuss in-depth some representative cases. The experimental results confirm that this AVESA framework effectively analyses epidemiological survey reports.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 91846303), and the Beijing Municipal Natural Science Foundation (No. 4212043).