Postprint: COVID-19 News Element Identification Using Transfer Learning Fusion in Multi-task Environments

Authors: Zhao Zibo, Wang Hao, Youhua Liu, Zhang Wei, Meng Town

Date: 2023-10-08T00:00:00+00:00

Abstract

[Objective/Significance] In the context of the COVID-19 pandemic, this study proposes a method for recognizing epidemic news elements that integrates transfer learning in a multi-task environment, to provide knowledge services for emergency events to the public. [Methodology/Process] First, news elements are identified via multi-task learning: temporal elements are recognized through rule-based methods; and a cross-domain element recognition model is constructed by integrating model transfer with deep learning methods. On this basis, association data for epidemic news elements is constructed, with the relationships between elements displayed via a knowledge graph. [Results/Conclusion] Experimental results show that the F1 scores for news element recognition (excluding drugs) all exceed 80%, indicating that the model integrating transfer learning achieves superior recognition performance; furthermore, the knowledge graph of association data can intuitively display the key elements and main content of news. In summary, the proposed method can effectively identify COVID-19 news elements, thereby helping news readers to accurately and efficiently obtain important information from news.

Full Text

Research on Identification of COVID-19 News Elements Based on Transfer Learning in a Multi-Task Environment

Zhao Zibo[1,2], Wang Hao[1,2], Liu Youhua[1], Zhang Wei[1,2], Meng Zhen[1,2]

[1] School of Information Management, Nanjing University, Nanjing 210023
[2] Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing 210023

Abstract:
[**Purpose/Significance**] Against the backdrop of the COVID-19 pandemic, this paper proposes a method for identifying COVID-19 news elements in a multi-task environment that integrates transfer learning, aiming to provide knowledge services for emergency events to the public.
[**Method/Process**] First, news elements are identified through multiple tasks: time elements are identified using rule-based methods, while a cross-domain element recognition model is constructed by integrating model transfer with deep learning approaches. On this basis, associated data of pandemic news elements is constructed, and the relationships between elements are displayed through knowledge mapping.
[**Result/Conclusion**] Experimental results show that the F1 scores for all news elements except drugs exceed 80%, indicating that the transfer learning model achieves favorable recognition performance. Moreover, the knowledge map of associated data can intuitively display key elements and main content of news. In summary, the proposed method can effectively identify COVID-19 news elements, thereby helping readers accurately and efficiently obtain important information from news.

Keywords: multi-task; transfer learning; COVID-19; news element identification; named entity recognition; cold start

The rapid proliferation of news reports has created significant psychological pressure and reading burden for the public. Therefore, it is necessary to quickly and accurately extract key elements from news reports to help the public obtain and understand main content, provide data support for further construction of pandemic news knowledge graphs [1], and lay the foundation for tasks such as automatic generation of pandemic news keywords [2] and proactive push of pandemic news [3].

News elements typically include four basic categories: time, person, location, and organization. COVID-19 news additionally involves medical elements such as disease names, symptoms, drug names, and diagnostic or treatment methods. Therefore, identifying COVID-19 news elements requires recognizing multiple categories across domains, involving multi-task and multi-process element identification.

Time elements exhibit strong regularity in expression and can be accurately identified through rule-based templates [4-5]. Thus, we adopt a rule-based approach for time element identification. For the three basic elements of person, location, and organization names, as well as the four medical elements of disease, symptom, drug, and method, we utilize named entity recognition (NER) methods based on existing deep learning models. However, as a novel type of emergency information resource, pandemic news faces a data cold-start problem due to the lack of annotated data for training NER models. To address this, we introduce transfer learning concepts and design a cross-domain entity recognition model.

Deep learning models, due to their strong learning capabilities, are prone to overfitting and require large-scale annotated data as training sets. Some domains suffer from data cold-start problems due to insufficient training data. To solve this issue, the concept of transfer learning [23] has emerged, which applies knowledge learned in a source domain to tasks in a different but related target domain, using annotated data from the source domain to train models applicable to the target domain. Based on the mature three-layer BERT-BiLSTM-CRF architecture in the NER field, we respectively utilize the MSRA dataset and a medical domain dataset to train transferable NER models and apply them to element identification in COVID-19 news. Finally, by constructing element association data based on co-occurrence frequency, we visually display the relationships between pandemic news elements in the form of a knowledge graph, thereby clearly and intuitively revealing the main content of pandemic news.

2 Recent Related Research

Transfer learning mainly includes instance-based, feature-based, and model-based approaches. Instance-based transfer learning expands the training set by adding source domain samples similar to target domain instances [24-25]. Feature-based transfer learning obtains and utilizes common feature representations between source and target domains to achieve representation-level transfer [26-27]. Model-based transfer learning transfers models and parameters trained on source domain data to the target domain [28-29]. Model-based transfer learning trains pre-trained models with strong generalization capabilities based on large-scale source domain data, enabling better adaptation to target domain data distributions and achieving superior transfer effects, making it widely applied in the NER field. M. Al-Smadi et al. constructed a multilingual universal sentence encoder based on transfer learning and applied it to entity recognition tasks in complex Arabic contexts [30]. Liu Yufei et al. transferred public domain source knowledge to the scientific domain to identify scientific terminology in patent documents [31]. Kong Xiangpeng et al. proposed a joint deep model based on transfer learning, training cross-language transfer models through shared network hidden layers and BP algorithm fine-tuning, effectively improving performance on Uyghur NER tasks [32].

While the above studies have achieved good entity recognition results, they have not examined model performance when using medical paper corpora as source domain training data. Considering that COVID-19 news is a type of real-time information resource for current emergency events lacking large-scale annotated corpora, we integrate model transfer with deep learning methods, using medical paper texts as the source domain dataset. Based on the effective BERT-BiLSTM-CRF three-layer architecture, we train entity recognition models and apply them to pandemic news element identification.

3 Methodology

3.1 Data Sources and Preprocessing

We selected The Paper's COVID-19 thematic series as the data source for pandemic news texts. As The Paper ranks among the top news media websites in China [33], its articles feature high quality with standardized vocabulary and syntax, making them suitable for news element extraction. Following the model transfer learning concept, we identified two source domain training datasets: (1) The Microsoft Research Asia (MSRA) dataset, a commonly used dataset for Chinese NER tasks containing over 27,000 sentences, used in this study to train the basic element recognition model COV19News-Base for identifying person, location, and organization elements; (2) The medical text dataset, derived from Chinese medical paper bibliographic data on COVID-related topics from CNKI, containing over 12,000 sentences after processing, used to train the medical element recognition model COV19News-Med for identifying disease, symptom, drug, and method elements. Source domain datasets are annotated in IOB format, where B indicates the beginning character of an entity, I indicates other characters within the entity, and O indicates non-entity characters (e.g., B-PER for person entity beginning, I-METHOD for non-beginning characters in method entities).

We employed a semi-supervised approach to obtain labeled medical text datasets. The specific process was: (1) Using the search query "SU='新冠' + '新型冠状病毒' + '武汉肺炎' + '2019-ncov' +'covid-19' " in CNKI's professional search function to retrieve papers in the medical and health science category published after "2020-02-01", batch downloading and saving 6,000 bibliographic records; (2) Extracting keyword fields from bibliographic data and manually annotating entity categories for keywords, obtaining 530 annotated keywords; (3) Using the Hownet synonym dictionary combined with manual supplementation to annotate synonyms of original words as the same category and add them to the keyword set, expanding the keyword set to 607 keywords; (4) Extracting all abstract fields from bibliographic data and using a maximum matching algorithm with the expanded keyword set to match sentences in abstract texts, generating a medical text corpus containing over 12,000 sentences with medical entities. This approach requires only manual annotation of a small number of keywords to match and obtain a large number of entity-containing sentences, significantly reducing manual annotation time costs.

3.2 Research Framework

To achieve automated identification and extraction of COVID-19 news elements, we designed the research framework shown in Figure 1 [Figure 1: see original paper]. (1) First, dataset preparation and preprocessing. Collect MSRA dataset, medical paper bibliographic data, and COVID-19 news text data, then manually annotate entity categories for keywords in medical paper bibliographic data and expand the keyword quantity. Subsequently, use the expanded keyword set to match sentences in the abstract collection to obtain a labeled medical text dataset. (2) Train transferable element recognition models based on source domain datasets. Using the BERT-BiLSTM-CRF three-layer architecture, train the basic element recognition model COV19News-Base based on the MSRA dataset to identify person, location, and organization elements, and the medical element recognition model COV19News-Med based on the medical text dataset to identify disease, symptom, drug, and method elements. Extract a certain proportion of samples from the original datasets as test sets to evaluate model performance. (3) Apply element recognition models to pandemic news text element identification. Manually annotate some sentences from news texts as the target domain test set to evaluate the transfer performance of models COV19News-Base and COV19News-Med for pandemic news element identification. (4) Finally, construct element association graphs based on news elements. Use the combination of COV19News-Base and COV19News-Med to extract elements from large volumes of pandemic news texts, combine with time elements extracted through rule-based methods, construct COVID-19 news element association data, and visualize the relationships between elements in the form of a knowledge graph to clearly and intuitively reveal the main content of pandemic news.

Based on this framework, we address three key issues: (1) Multi-category element identification. Multiple element recognition tasks are divided, and rule-based and NER methods are used to identify basic elements, medical elements, and time elements in pandemic news. (2) Data cold-start problem. Model transfer learning is introduced, using sufficient annotated data from source domains to train transferable NER models and apply them to element identification in pandemic news, thereby solving the problem of insufficient annotated data in the target domain. (3) Utilization of pandemic news elements. The proposed element identification method is applied to large volumes of unlabeled pandemic news texts, and the identified elements and their co-occurrence relationships are stored as pandemic news element association data. Based on this, element association graphs are constructed to visually display relationships between elements, thereby revealing the main content of pandemic news.

3.3 COVID-19 News Element Classification

We aim to automatically identify and extract eight categories of pandemic news elements, with names and examples shown in Table 1 . Among them, time, person, location, and organization are basic elements describing news content. Additionally, COVID-19 news often contains medical elements such as disease names, symptoms, drug names, and diagnostic or treatment method names. For specific medical element categories, we can draw on previous research experience. In the 2019 China Conference on Knowledge Graph and Semantic Computing (CCKS) medical NER task, medical entities were divided into six categories: disease and diagnosis, examination, test, surgery, drug, and anatomical part [20]. The 2017 CCKS defined four medical entity categories: body part, symptom and sign, examination and test, and disease and diagnosis [34]. Zhao Qing et al. and Xia Guanghui et al. categorized medical entities into disease, symptom, examination, and treatment [35-36]. Summarizing these studies, medical entities include five categories: disease name, symptom/sign, drug, examination/treatment method, and body part. However, body part entities in news contexts often have meanings beyond disease locations (e.g., “hand” in “shake hands and make peace,” “mouth” in “pay lip service”), which do not refer to disease locations and are not key elements for extracting news points. In conclusion, we determined to identify disease, symptom, drug, and method as medical elements.

We identify various pandemic news elements through multiple tasks. For the seven element categories other than time, we adopt NER methods based on the BERT-BiLSTM-CRF model, training separate basic element recognition and medical element recognition models. For time elements, we use rule-based identification by constructing regular expression templates to match and extract time elements from news texts. The regular expression template for matching time elements is shown in formula (1):

pattern=[
r“4̣ 年 1,2̣ 月 1,2̣[日 | 号]”,
r“1,2̣ 月 1,2̣[日 | 号]””,
...
]

3.4 COVID-19 News Model Training Based on Transfer Learning

Due to the lack of annotated data for training NER models in the pandemic news domain, we employ a model training method integrating transfer learning, training models COV19News-Base and COV19News-Med based on the MSRA dataset and medical text dataset respectively, and applying these models to identify various elements in pandemic news texts. To evaluate different model performances, we divide both the MSRA dataset and medical text dataset into training and test sets.

Before model training, we statistically analyzed entity quantities in source domain training sets, source domain test sets, and target domain test sets, with results shown in Table 2 . The source domain dataset for model COV19News-Base is the MSRA dataset, while the source domain dataset for model COV19News-Med is the medical text dataset; the target domain test set for both models consists of sentences extracted from news texts. Table 2 reveals uneven entity distribution in source domain datasets: location entities significantly outnumber person and organization entities in the MSRA dataset, while disease entities far exceed other three entity types in the medical text dataset. This is because most keywords from medical papers are disease entities, primarily including numerous aliases for COVID-19, causing uneven distribution of matched entity quantities. From the target domain test set perspective, person, location, and organization entities are relatively evenly distributed, while disease entities remain the most frequent among medical entities, consistent with characteristics of COVID-19 news (reports containing many references and aliases for COVID-19). Whether this uneven entity distribution affects model performance requires experimental verification. Additionally, the medical text dataset is relatively smaller than the MSRA dataset, resulting in fewer entities for training, which may impact model effectiveness and warrants further investigation.

Based on the BERT-BiLSTM-CRF model, we trained models COV19News-Base and COV19News-Med using the above training data. BERT employs a multi-layer bidirectional Transformer [37] encoder structure that can capture long-range contextual semantic features to obtain precise text vectors. BiLSTM uses dual reverse-order LSTM networks to fully learn bidirectional semantic relationships between vectors. CRF can output globally optimal label sequences according to sequence labeling constraint rules. Therefore, the BERT-BiLSTM-CRF model achieves optimal learning effects at the representation, network, and output layers, making it suitable for COV19News model training. After training, we evaluated model performance on source domain and target domain test sets, with results presented in the Experimental Results and Analysis section.

3.5 Knowledge Graph Construction for Pandemic News Elements

After identifying and extracting pandemic news elements using the above models, we further construct a knowledge graph for pandemic news elements to visually display relationships between elements.

Considering that associations exist between pandemic news elements and that these associations can reveal the main events of news stories, mining element association relationships helps infer the main content of pandemic news and is significant for readers' comprehension. We first divide entire news texts into sentence collections, then record elements appearing in the same sentence as co-occurring once, calculate pairwise element co-occurrence frequencies, and save them in the format "Element A-Element B-Co-occurrence Frequency" as data files serving as pandemic news element association data. This association data describes relationships between elements and their strength, providing data support for constructing pandemic news element knowledge graphs.

The pandemic news element knowledge graph can clearly and intuitively display element associations and their strengths, helping readers locate key elements in news and infer main content. Therefore, based on news element association data, we construct an association knowledge graph using elements as nodes and co-occurrence frequencies between two elements as edge weights between nodes.

We used the network analysis software Gephi to draw the pandemic news element association knowledge graph, as shown in Figure 2 [Figure 2: see original paper]. Figure 2 shows that key elements with dense connections to other elements are prominently displayed, and readers can connect various elements to infer the main news content based on element associations.

4 Experiments and Analysis

4.1 Experimental Environment and Model Parameter Settings

Model training, testing, and transfer were conducted on a personal computer with an NVDIA GeForce RTX 2060 GPU (6GB memory), 16GB RAM, and Windows 10 operating system. The model runtime environment was Python 3.5 + TensorFlow 1.12 GPU version, with CUDA version 10.2.

Partial parameters of the BERT-BiLSTM-CRF model are shown in Table 3 .

Table 3. Model Parameter Settings

Parameter	Description
Batch Size	Number of samples (characters) fed into the training model at once
Segment Embedding Size	Vector dimension representing BERT sentence embedding
Char Embedding Size	Vector dimension representing BERT character embedding
Dropout Rate	Forgetting rate of BiLSTM network
Learning Rate	Learning rate of BiLSTM network
Optimizer	Optimizer type used for network forward propagation
Max Epoch	Maximum number of training iterations

4.2 Testing and Transfer of Model COV19News-Base

We evaluate model performance using Precision (P), Recall (R), and their harmonic mean F1-measure (F1). For entities typically containing multiple characters, an entity is considered correctly identified only when the model's output label sequence exactly matches the original annotation sequence; otherwise, it is considered incorrect. In subsequent experiments, OP, OR, and OF1 represent P, R, and F1 values in the source domain, while TP, TR, and TF1 represent P, R, and F1 values in the target domain.

Training model COV19News-Base on the MSRA dataset, its performance on source domain and target domain test sets is shown in Figure 3 [Figure 3: see original paper]. The results indicate: (1) Due to consistent entity distribution characteristics between training and test sets in the same domain, the model demonstrates superior identification performance on the source domain test set, with F1 values for all three entity types exceeding 90%. (2) After transferring to the target domain, identification performance for all three entity types declined to varying degrees, but F1 values remained above 80%. Considering the distribution differences between pandemic news texts and the MSRA dataset, this slight performance decline after transfer is expected. (3) Comparing identification performance across the three entity types, person entities achieve the best results, followed by location entities, while organization entities show the poorest performance. The average length of location and organization entities is typically greater than that of person entities, and their identification difficulty is correspondingly higher, leading to performance differences across entity types. (4) Although location entities appear more frequently than other two entity types in the source dataset, their identification performance is not superior, indicating that uneven entity distribution in the training set does not affect model performance.

4.3 Testing and Transfer of Model COV19News-Med

Replicating the base model parameters, we trained model COV19News-Med on the medical text dataset. Its performance on source domain and target domain test sets is shown in Figure 4 [Figure 4: see original paper]. The findings reveal: (1) The model still demonstrates superior performance on the source domain test set, with F1 values for all four medical entity types exceeding 90%, indicating that the BERT-BiLSTM-CRF framework possesses strong representation and learning capabilities, maintaining good fitting effects for data from different domains. (2) Although the medical text dataset is smaller than the MSRA dataset, its performance on the source domain test set is not inferior, suggesting that with sufficient data scale, relatively small sample sizes can still yield good training results without affecting model effectiveness. (3) After transferring to the target domain, identification performance for various entity types declined to varying degrees. However, except for drug entities, F1 values for the other three entity types remained above 80%, which is acceptable. The performance decline is primarily due to poor recall rates for various entities in the target domain test set, likely because of substantial distribution differences in medical entities between medical paper texts and pandemic news texts, resulting in suboptimal generalization after transfer and making some entities present in the target domain but not learned by the model difficult to identify. Nevertheless, the transferred model still maintains high identification precision. (4) In the source domain dataset, disease entities far outnumber the other three entity types, and disease entities also show the best performance in both source and target domain test sets. However, the F1 value gap with symptom and method entities in the target domain test set is no longer significant. This suggests that while extremely uneven entity distribution may positively affect identification of certain entity types in the source domain, it does not necessarily significantly impact their performance in the target domain, which remains related to entity distribution characteristics in the target domain.

The above experimental results demonstrate that NER models trained using transfer learning methods achieve favorable identification performance for target domain pandemic news elements. To showcase the identification effectiveness of our proposed method, we randomly selected multiple sentences containing multiple element types from pandemic news texts and used models COV19News-Base and COV19News-Med to identify their elements. Based on regular expression templates for time element patterns, we matched and identified time elements in sentences, then aggregated results from multiple tasks. Partial results are shown in Table 4 .

Table 4. Examples of Pandemic News Element Identification Results

Original Sentence	Identified Elements
That was January 21, 2020. The novel coronavirus pneumonia epidemic was spreading from Wuhan to the whole country. Weng Qiuqiu was in Qichun County, Huanggang, Hubei, only a hundred miles from Wuhan. Huanggang was the most severely affected area outside Wuhan.	{January 21, 2020, novel coronavirus infection, pneumonia, Wuhan, Weng Qiuqiu, Hubei, Huanggang, Qichun County}
Since January 26, with symptoms such as fatigue and cough, Wu Juan's father had been waiting for a definitive diagnosis—without confirmed novel coronavirus pneumonia (hereinafter referred to as "COVID-19"), he could not be admitted to the hospital.	{January 26, fatigue, cough, Wu Juan, novel coronavirus infection, pneumonia, COVID-19}
Xu Shunqing, Deputy Dean of the School of Public Health at Huazhong University of Science and Technology, stated at a press conference in Hubei Province on the evening of February 9 that the novel coronavirus mainly attacks the lungs, so nucleic acid testing has a certain false negative rate, meaning some patients are not detected, resulting in missed diagnoses, which may cause some infection sources not to be truly identified, posing a risk of expansion.	{School of Public Health, Huazhong University of Science and Technology, Xu Shunqing, February 9, Hubei Province, novel coronavirus, nucleic acid testing}

Original Sentence	Identified Elements
When Austria notified on March 4 that an Austrian man on the cruise was infected with the novel coronavirus, the "Opera" cruise ship was docked at Piraeus Port in Athens. For safety reasons, the "Opera" cruise ship requested all passengers who had disembarked at Piraeus Port to return to the ship as soon as possible.	{Austria, March 4, novel coronavirus, "Opera" cruise ship, Athens, Piraeus Port}
Li Wenliang, male, 35 years old, ophthalmologist at Wuhan Central Hospital. Li Wenliang took the college entrance examination in 2004. After graduating from the seven-year clinical medicine program at Wuhan University, he first worked in Xiamen for three years, returned to Wuhan in 2014, and has been working at Wuhan Central Hospital ever since.	{Li Wenliang, Wuhan Central Hospital, Wuhan University, Xiamen, Wuhan}

4.4 Construction of COVID-19 News Element Knowledge Graph

Based on the above pandemic news element identification method, we extracted news elements and constructed element association data to build an associated knowledge graph for COVID-19 news elements. Taking a news article titled "Family Account | The Last 12 Days of a 'Severe Pneumonia' Patient" as an example, we constructed its element knowledge graph, as shown in Figure 5 [Figure 5: see original paper].

Figure 5 shows that this news article mainly involves time, person, location, organization, and disease elements, with "Weng Qiuqiu," "Wuhan," and "pneumonia" as key elements. Based on element associations, we can infer that the main content of this news is about Weng Qiuqiu, a citizen of Huanggang, who contracted COVID-19 and received treatment at Huanggang Traditional Chinese Medicine Hospital. This demonstrates that the pandemic news element association knowledge graph can effectively help readers identify news priorities and infer main content, showing potential as a novel knowledge service for COVID-19 emergency events.

5 Conclusion

We propose a pandemic news element identification method that integrates transfer learning and deep learning technologies in a multi-task environment, providing a feasible service solution for public information acquisition during emergency events. First, combining NER with rule-based methods, we identify multi-category news elements through multiple tasks. Simultaneously, to solve the data cold-start problem in the pandemic news domain, we adopt a model transfer solution to obtain cross-domain element recognition models with good identification performance. Finally, we apply the identification scheme to large volumes of pandemic news texts, constructing an element association knowledge graph based on identified news elements to help readers intuitively and quickly discover key elements and main content.

Through comparative analysis of model testing and transfer performance, we draw the following conclusions: (1) The three-layer BERT-BiLSTM-CRF architecture is suitable for NER tasks in different domains, with F1 values for all entity types in source domains exceeding 90%; (2) After transferring models from source to target domains, identification performance declines but remains within acceptable ranges, with F1 values for most entity types above 80%; (3) If entity distribution in source domain training data is extremely unbalanced, it may lead to over-learning of certain entity types, with identification performance for those types far exceeding others in the source domain, but whether this affects target domain identification performance requires further research.

In summary, our proposed transfer learning-based element identification method achieves favorable identification performance for COVID-19 news elements. However, this study still has issues such as lower identification rates for some entity categories. Future research will focus on combining instance transfer with model transfer to make entity distributions between training and target domains more similar, thereby improving model identification performance in target domains.

References

[1] Wang Yan, Hao Xinghua, Xue Peng. Construction and Analysis of Linked Data Knowledge Graph Based on Co-word Analysis and Social Network Analysis[J]. Digital Communication World, 2020(6):148-150.
[2] Tao Jie. Keyword Extraction Based on News Text[D]. Wuhan: Central China Normal University, 2019.
[3] Tao Tianyi, Wang Qingqin, Fu Yuwei, et al. Financial News Personalized Recommendation Algorithm Based on Knowledge Graph[J/OL]. Computer Engineering, 2020: 1-10 [2020-09-12]. https://doi.org/10.19678/j.issn.1000-3428.0057446.
[4] Pei Tao, Guo Sihui, Yuan Yecheng, et al. Research on Structuring Network Text Big Data for Public Safety Events[J]. Journal of Geo-Information Science, 2019, 21(1):2-13.

[5] Ji Leijing. Research on Semantic Detection Method for Geographic Information Change in Web Pages[D]. Nanjing: Nanjing Normal University, 2013.
[6] Fu Kai. Research on Web News Text Information Extraction and Visualization[D]. Jinan: Shandong University of Finance and Economics, 2017.
[7] Krstev C, Obradovic I, Utvic M, et al. A System for Named Entity Recognition Based on Local Grammars[J]. Journal of Logic and Computation, 2014, 24(2):473-489.
[8] Yang Jianlin, Wang Wenlong. Research on Extraction of Public Health Emergencies[J]. Information Studies: Theory & Application, 2016, 39(4):51-59.
[9] Kucuk D, Yazici A. A Hybrid Named Entity Recognizer for Turkish[J]. Expert Systems with Applications, 2012, 39(3):2733-2742.
[10] Seker G A, Eryigit G. Extending a CRF-based Named Entity Recognition Model for Turkish Well Formed Text and User Generated Content[J]. Semantic Web, 2017, 8(5):625-638.
[11] Wu Weicheng. Research on Temporal Phrase Extraction Based on Terrorist Attack Event Corpus[D]. Nanjing: Nanjing University, 2016.
[12] Chasin R, Woodward D, Witmer J, et al. Extracting and Displaying Temporal and Geospatial Entities from Articles on Historical Events[J]. Computer Journal, 2014, 57(3):403-426.
[13] Li Yuchao. Research on News Event Place Name Entity Recognition and Map Linking Technology[D]. Chengdu: University of Electronic Science and Technology of China, 2020.
[14] Wichmann P, Brintrup A, Baker S, et al. Extracting Supply Chain Maps from News Articles Using Deep Neural Networks[J]. International Journal of Production Research, 2020, 58(17):5320-5336.
[15] Xu J G, Guo L X, Jiang J, et al. A Deep Learning Methodology for Automatic Extraction and Discovery of Technical Intelligence[J]. Technological Forecasting and Social Change, 2019, 146:339-351.
[16] Wang Hao, Deng Sanhong, Zhu Liping, et al. Research on the Intelligence Value and Utilization of Government Data in Big Data Environment: Taking Customs Declaration Commodity Classification Risk Avoidance as an Example[J]. Technology Intelligence Research, 2020, 2(4):74-89.
[17] Dong X S, Chowdhury S, Qian L J, et al. Deep Learning for Named Entity Recognition on Chinese Electronic Medical Records: Combining Deep Transfer Learning with Multitask Bi-directional LSTM RNN[J]. PLOS One, 2019, 14(5):1-15.
[18] Xiao Lianjie, Meng Tao, Wang Wei, et al. Research on Intelligence Analysis Method Identification Based on Deep Learning: Taking Security Intelligence as an Example[J]. Data Analysis and Knowledge Discovery, 2019, 3(10):20-28.
[19] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[EB/OL]. [2020-09-12]. https://arxiv.org/abs/1810.04805.
[20] Li Lingfang, Yang Jiaqi, Li Baoshan, et al. Chinese Electronic Medical Record Named Entity Recognition Based on BERT[J]. Journal of Inner Mongolia University of Science and Technology, 2020, 39(1):71-77.
[21] Wu Jun, Cheng Yao, Hao Han, et al. Research on Chinese Professional

Term Extraction Based on BERT-Embedded BiLSTM-CRF Model[J]. Journal of the China Society for Scientific and Technical Information, 2020, 39(4):409-418.
[22] Liu Zhongbao, Dang Jianfei, Zhang Zhijian. Research on Automatic Extraction of Historical Events and Construction of Event Logic Graph in "Records of the Grand Historian"[J]. Library and Information Service, 2020, 64(11):116-124.
[23] Yosinski J, Clune J, Bengio Y, et al. How Transferable are Features in Deep Neural Networks?[EB/OL]. [2020-09-12]. https://arxiv.org/abs/1411.1792.
[24] Chen Meishan, Xia Chenxi. Named Entity Recognition for Liver Cancer Patients' Online Questions: A Transfer Learning-Based Method[J]. Data Analysis and Knowledge Discovery, 2019, 3(12):61-69.
[25] Li Haohao. Research and Application of Instance-Based Transfer Learning Technology[D]. Wuhan: Wuhan University, 2018.
[26] Chen Wenjun, Yang Jiajia. Research on Cross-Domain Recommendation Based on Shared Knowledge Transfer Learning[J]. Information Science, 2020, 38(6):126-132.
[27] Gligic L, Kormilitzin A, Goldberg P, et al. Named Entity Recognition in Electronic Health Records Using Transfer Learning Bootstrapped Neural Networks[J]. Neural Networks, 2020, 121:132-139.
[28] Kung H K, Hsieh C M, Ho C Y, et al. Data-Augmented Hybrid Named Entity Recognition for Disaster Management by Transfer Learning[J]. Applied Sciences-Basel, 2020, 10(12):1-17.
[29] Shao Mingrui, Ma Denghao, Chen Yueguo, et al. Research on FAQ Question-Answering Model Based on Community Q&A Data Transfer Learning[J]. Journal of East China Normal University (Natural Science), 2019(5):74-84.
[30] Al-Smadi M, Al-Zboon S, Jararweh Y, et al. Transfer Learning for Arabic Named Entity Recognition with Deep Neural Networks[J]. IEEE Access, 2020, 8:37736-37745.
[31] Liu Yufei, Yin Li, Zhang Kai, et al. Technical Term Recognition Based on Deep Transfer Learning: Taking the CNC System Domain as an Example[J]. Journal of Intelligence, 2019, 38(10):168-175.
[32] Kong Xiangpeng, Wushour・Silamu, Yang Qimeng, et al. Uyghur Named Entity Recognition Based on Transfer Learning[J]. Journal of Northeast Normal University (Natural Science Edition), 2020, 52(2):58-65.
[33] Webmaster Home. News Media Website Ranking[EB/OL]. [2020-09-30]. https://top.chinaz.com/hangye/index_{news}.html.
[34] Li Fei, Zhu Yanhui, Wang Tianji, et al. Research on Electronic Medical Record Named Entity Recognition Based on Medical Categories[J]. Journal of Hunan University of Technology, 2018, 32(4):61-66.
[35] Zhao Qing, Wang Dan, Xu Shushi, et al. A Weakly Supervised Chinese Medical Entity Recognition Method Based on RNN[J/OL]. Journal of Harbin Engineering University, 2020:1-10 [2020-09-12]. http://kns.cnki.net/kcms/detail/23.1390.U.20200330.1522.002.html
[36] Xia Guanghui, Li Junlian, Xing Baokun, et al. Research on Medical Diagnosis and Treatment Named Entity Recognition Based on Chinese Case Report

Literature[J]. Journal of Medical Informatics, 2019, 40(6):54-59.
[37] Vaswani A, Shazeer N, Parmar N, et al. Attention is All You Need[EB/OL]. [2020-09-12]. https://arxiv.org/abs/1706.03762.

Author Contributions

Zhao Zibo: Conducted experiments and wrote the initial manuscript.
Wang Hao: Guided research ideas, reviewed paper content, and provided revision suggestions.
Liu Youhua: Organized experimental results, reviewed abnormal data indicators, and proposed improvements.
Zhang Wei: Provided guidance on visualization methods and tools, and participated in revisions.
Meng Zhen: Responsible for final manuscript revisions.

Note: Figure translations are in progress. See original paper for figures.