# A Study on Precise Identification of Domain Information Using Transfer Learning (Postprint)

**Authors:** Lu Quan, Zhitong Hao, Chen Jing, Chen Shi, Zhu Anqi

**Date:** 2023-04-01T16:02:47+00:00

## Abstract

[Purpose/Significance] This study migrates unsupervised learning results from large-scale Internet data to target domains, addressing the challenge of improving information recognition performance in target domains with limited training samples. [Method/Process] We utilize a RoBERTa model pretrained on Chinese Wikipedia and other data for transfer learning; after mapping the learned representations to the target domain, we employ DPCNN for aggregation and distillation, then fine-tune the model with partially labeled data to accomplish accurate domain information identification. [Results/Conclusion] When compared across 10 domains with both non-transfer-learning models and the classical TextCNN model, the proposed model significantly outperforms the baselines, achieving absolute improvements of 4.15% and 3.43% in precision, 4.55% and 3.44% in recall, and 4.52% and 3.44% in F1-score, respectively, demonstrating that transfer learning utilizing large-scale web data can significantly enhance information recognition performance in target domains.

## Full Text

## Preamble

**Volume 65, Issue 5, March 2021**

**Exploring the Use of Transfer Learning for Accurate Domain Information Identification**

Lu Quan[1],[2], Hao Zhitong[1], Chen Jing[3], Chen Shi[1], Zhu Anqi[1]

[1]Center for Studies of Information Resources, Wuhan University, Wuhan 430072
[2]Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources, Shenzhen 518034
[3]School of Information Management, Central China Normal University, Wuhan 430079

**Abstract:** [Purpose/Significance] This study transfers unsupervised learning results from large-scale internet data to target domains to address the challenge of limited learning samples that hinder information identification performance. [Method/Process] We employ the RoBERTa model pretrained on Chinese Wikipedia and other corpora for transfer learning, map the learning results to the target domain, aggregate and condense them using DPCNN, and then fine-tune the model with partially annotated data to achieve accurate domain information identification. [Results/Conclusion] Compared with non-transfer learning models and the classic TextCNN model across ten domains, our proposed model demonstrates substantial improvements, achieving absolute increases of 4.15% and 3.43% in precision, 4.55% and 3.44% in recall, and 4.52% and 3.44% in F1-score on average. These results indicate that transfer learning from web-scale data can significantly enhance information identification effectiveness in target domains.

**Keywords:** transfer learning; information identification; RoBERTa
**Classification Number:** TP391.1
**DOI:** 10.13266/j.issn.0252-3116.2021.05.011

Domain information identification represents a crucial research direction in natural language processing (NLP), attracting sustained attention from scholars in computer science, linguistics, and related fields. Its objective is to separate domain-specific information from text collections across different domains [1]. Conventional approaches typically involve feature engineering to construct a series of features followed by machine learning processing, or directly training deep neural networks such as CNN (Convolutional Neural Network) and RNN (Recurrent Neural Network) to learn hidden features for information identification. These methods essentially build rules from observed data to infer patterns in unobserved data [2], which means that sufficient labeled data is required for the model to achieve accurate inference. To ensure model performance, collecting large-scale labeled data for specific tasks has become a rigid requirement, posing significant challenges for many application scenarios of supervised learning. These datasets often require manual annotation, which is costly, time-consuming, and error-prone. Vertical domain information identification faces particular difficulties, as limited domain data and annotation challenges become key factors constraining model performance.

In contrast, with the rise of Web 2.0, users have become important creators of virtual community content, continuously generating massive amounts of data on the internet, including high-quality sources like Wikipedia. To leverage the advantages of internet-scale data to compensate for insufficient domain data, we utilize the RoBERTa pretrained model for transfer learning. To condense the highly dispersed output results from transfer learning for domain-specific information, we employ DPCNN (Deep Pyramid Convolutional Neural Networks) to aggregate the transfer results and discriminate domain information. Finally, we conduct experimental validation across different domains to evaluate the effectiveness of our constructed model.

## 2 Related Research

### 2.1 Domain Information Identification

Domain information identification involves comparing the form of an object's motion state and its changes, or certain characteristic parameters of this form, with the form or characteristic parameters of a "domain template" with specific attributes, and determining the domain category to which the information belongs based on the differences in their matching [3]. Broadly defined, domain information identification includes text, image, and speech recognition, while narrowly defined, it refers only to text-based information identification [4]. This paper adopts the narrow definition.

With the rapid development of information technology, electronic texts across various domains have grown exponentially and are often disorderly distributed within internet communities, creating difficulties for precise domain research. Manual processing of such volumes is impractical, necessitating technical approaches for automated domain information identification. Scholars have proposed numerous classical methods through in-depth research, which can be summarized into two categories: statistical and machine learning-based methods, and deep learning-based methods.

Statistical and machine learning-based methods generally consist of two steps: first, feature engineering based on character and word statistics [5], followed by information identification using machine learning algorithms on the selected features. For example, Liao Liefa et al. [6] utilized LDA (Latent Dirichlet Allocation) to extract topic features and employed KNN (K-Nearest Neighbor) classification to identify rare earth domain patent documents, achieving good accuracy. Yang Tengfei et al. [7] used co-word analysis and SVM (Support Vector Machine) to identify typhoon disaster information within Weibo posts. While these statistical and machine learning methods advanced domain information identification for a period and yielded excellent results, the separation of feature extraction and identification algorithms made performance highly dependent on feature engineering. Feature quality directly determined the model's ceiling, imposing high demands on human expertise. Consequently, with the rise of neural networks, scholars gradually shifted their focus to deep learning models that require no manual feature construction.

Deep learning methods simulate human brain neurons, typically vectorizing characters and words from target and non-target domains and feeding them into layered neural networks. They adjust connection strengths between neurons using loss functions to distinguish hidden features and achieve domain information identification. Y. Kim [8] was among the first to apply CNN from computer vision to text information, proposing the TextCNN model, which immediately set new records on multiple open datasets including MR (Movie Review) upon release. Subsequently, Huang Tao [9] used this model to identify news information across different domains. Later, a series of excellent models emerged, including RCNN [10], fastText [11], and DPCNN [12]. DPCNN, in particular,

can capture longer text dependencies without significantly increasing computational cost and has profoundly influenced the NLP field. Currently, deep learning-based methods represent the mainstream approach for domain information identification, widely applied in both academia and industry. However, deep learning methods suffer from severe cold-start problems, requiring large amounts of domain-labeled data for model parameter training.

### 2.2 Transfer Learning

Transfer learning is a machine learning approach that leverages knowledge from a relatively mature domain (source domain) to solve problems in a related but immature domain (target domain). It effectively relaxes the two prerequisites of traditional machine learning: "the learning process requires large labeled datasets" and "the test set and training set must satisfy the same distribution assumption" [13]. The emergence of transfer learning provides solutions for target domains with limited or no annotations.

Transfer learning was first applied in computer vision. B. Zhou et al. [14] pretrained on ImageNet and Places datasets and compared transfer learning results with small-scale dataset results, powerfully demonstrating that transfer learning can achieve better performance. In NLP, transfer learning started relatively later. T. Mikolov et al. [15] proposed word2vec, which uses large-scale corpora to learn word semantics only for the model's first layer, which can be directly used as a word embedding layer for other models. This approach had significant impact, but target tasks still required training from scratch. Until 2017, when Google proposed the Transformer architecture [16], which holds milestone significance in NLP, subsequent transfer learning pretrained models have almost all been based on Transformer. In 2018, the Wikipedia-based pretrained models GPT [17] and BERT [18] emerged, nearly 刷新 ing all NLP task leaderboards. Later, scholars proposed improved models based on BERT, including RoBERTa [19] and ALBERT [20], with RoBERTa achieving better results through improved pretraining tasks and larger batch sizes. This paper uses the RoBERTa model for transfer learning.

### 2.3 Domain Information Identification from a Transfer Learning Perspective

From a transfer learning perspective, domain information identification primarily involves two problems: how to perform transfer learning, and how to use the output results of the transfer learning model.

For the first problem, with the emergence of pretrained models, researchers in academia and industry have gradually reached a consensus on using pretrained models for transfer learning. As long as sufficiently rich corpora are invested in pretraining, this completely avoids the need to train transfer learning models from scratch—a process that often consumes months. Currently, a few foreign scholars have begun using pretrained models for transfer learning to complete do-

main information identification. For example, N. Houlsby et al. [21] conducted information identification across aviation, economics, and natural disaster domains on 17 public datasets, finding that using pretrained BERT improved identification performance by at least 0.4%. T. Sharma et al. [22] achieved better identification results for food information based on the RoBERTa pretrained model. However, Chinese transfer learning pretrained models emerged relatively late, so no domestic scholars have been observed using them for domain information identification.

For the second problem, most existing research simply applies the pretrained model's output through a slightly modified Softmax layer directly to specific domain information identification tasks. While this simple usage can achieve good results, the pretrained model for transfer learning is trained through unsupervised learning on massive data and is not targeted at specific problems. Its output is a highly dispersed long sequence. Relying entirely on pretrained output without more refined adjustments can easily waste the results learned through transfer.

To address these issues, we designed a solution that "uses the RoBERTa model pretrained on large-scale Chinese corpora for transfer learning, and aggregates and condenses the transfer learning results through DPCNN for domain information identification," aiming to achieve better identification performance. In summary, transfer learning remains an emerging research field, and domain information identification based on transfer learning still has considerable room for improvement, warranting further in-depth research from scholars.

## 3 Construction of a Domain Information Identification Model Based on Transfer Learning

To leverage the advantages of web-scale data, we use RoBERTa, which set multiple NLP task records in 2019, as the pretrained model for transfer learning. Since the transfer learning model's output is a highly dispersed long sequence, we employ the DPCNN model to capture long-distance dependencies in the output sequence and aggregate information for domain information judgment. The final model architecture is shown in Figure 1 [Figure 1: see original paper].

First, the dataset undergoes embedded representation (Embedding) according to a three-layer structure of character vectors, segment vectors, and position vectors, which is then passed into the RoBERTa model pretrained on Chinese Wikipedia and other data. Next, DPCNN is used for domain information identification, and the pretrained model parameters are fine-tuned based on feedback from annotated data. Finally, identification results for specified domain information are obtained on a test set containing information from numerous domains, enabling model evaluation.

### 3.1 Text Representation

The model's input consists of three combined layers: character vectors, segment vectors, and position vectors, as shown in Figure 2 [Figure 2: see original paper]. The first layer, character vectors, represents the embedding of each token in the vocabulary. The second layer, segment vectors, distinguishes different sentences in the input text; segment vectors are identical for tokens within the same sentence and can be set to all zeros for single-sentence inputs. The third layer, position vectors, records the relative and absolute positions of each character. The [CLS] and [SEP] tokens in the figure represent the input text marker and sentence separator, respectively.

The position vectors can be calculated according to formula (1), where pos represents the absolute position of a character/word, i denotes the position in the embedding dimension, and d represents the vector dimension.

$$PE(pos, 2i) = sin(pos/10000^{2i/d})$$
$$PE(pos, 2i+1) = sin(pos/10000^{2i+1/d})$$

The model then predicts masked characters. Although only a small portion of the input text is predicted, this does not affect the model's language understanding capability due to the filling effect of large-scale corpora.

### 3.2 Transfer Learning Based on RoBERTa

RoBERTa is a transfer learning model for NLP tasks released by Facebook in 2019, achieving state-of-the-art results on the GLUE, SQuAD, and RACE benchmarks. During pretraining, the model adopts BERT's masked language model mechanism but differs from BERT's static masking by dynamically randomizing the masking for each input sequence. It then predicts masked words based on context, as illustrated in Figure 3 [Figure 3: see original paper]. Specifically, 15% of each input sequence is randomly selected for special processing: 80% are replaced with [Mask], 10% are randomly replaced with other tokens, and 10% remain unchanged.

Notably, whole-word masking is employed here: if a masked character is part of a word, all characters in that word are masked. This enhances the model's ability to handle complex problems. For example, in Figure 3, predicting "习" given "学" is undoubtedly easier than directly predicting "学习".

The internal structure of the RoBERTa model is shown in Figure 4 [Figure 4: see original paper], where each "Trm" represents a Transformer Encoder. It is evident that during layer-by-layer progression, each "Trm" uses attention mechanisms to connect with all "Trm" units in the previous layer. This structure utilizes information in a comprehensive manner, unlike traditional LSTM (Long

Short-Term Memory) or Bi-LSTM (Bidirectional LSTM) that can only pass information unidirectionally or bidirectionally.

Moreover, this structure is highly parallelized: nodes in each sequence are generated simultaneously, with individual nodes not depending on previous or subsequent computation results. Consequently, the RoBERTa model based on this structure can process more corpora within the same time frame, which is key to its effectiveness as a transfer learning model.

### 3.3 DPCNN-Guided Fine-Tuning and Identification

Transfer learning produces a long sequence that cannot be directly applied to domain information identification and must be further fine-tuned according to task requirements. To capture dependencies between distant nodes in this long sequence, we use DPCNN, proposed by Tencent AI Lab in 2017, to aggregate domain information and make judgments, then feed the results back to the transfer learning model for parameter fine-tuning. This process is shown in Figure 5 [Figure 5: see original paper].

DPCNN continuously feeds the RoBERTa output embeddings into two convolutional layers followed by 1/2 pooling, then repeats this process. To avoid gradient vanishing, residual connections are used between inputs and outputs during repetition. Specifically, max pooling with size 3 and stride 2 is performed after each convolutional block. This pooling strategy halves the size of each document's internal representation. With a fixed number of feature maps, the effective coverage of convolutional kernels doubles after each downsampling period. Consequently, words within twice the distance become associated after a downsampling cycle, meaning DPCNN efficiently captures information from greater distances and ultimately utilizes global information.

Additionally, when downsampling with stride 2, each convolutional layer's computation time is halved (data size reduced by half), forming a "pyramid". Thus, total computation time is bounded by a constant that is twice the computation time of the bottom layer, giving DPCNN computational efficiency advantages.

As shown more intuitively in Figure 6 [Figure 6: see original paper], after a series of deep convolutions and pooling, the original input sequence is highly compressed, equivalent to recording and aggregating long-distance information. Each repetition expansion allows upper-layer nodes to detect information from greater distances while halving the sequence length, enabling comprehensive information utilization for domain content identification and improving information usage efficiency.

During final fine-tuning based on domain information annotation, the paper uses cross-entropy loss function and Adam optimizer to adjust the entire network's weights. The cross-entropy loss is shown in formula (2), where Y and X represent the label and processed variable matrices, respectively, and h represents the sigmoid activation function.

$$J(\theta) = -Y^T \log h_\theta(X) - (E - Y)^T \log(E - h_\theta(X))$$

The final model then identifies domain information on the test set and completes model evaluation.

## 4 Experiments and Results Analysis

### 4.1 Dataset

This experiment uses the THUCNews dataset, released by Tsinghua University's Natural Language Processing Laboratory and reorganized by Hu Wenxing. The dataset was collected from Sina.com and consists of text data from ten domains: finance, real estate, stock market, education, technology, society, current affairs, sports, games, and entertainment. Each domain contains 20,000 data entries, totaling 200,000 entries. The dataset was partitioned for use according to a training set:validation set:test set ratio of 18:1:1.

The RoBERTa pretrained model [24] used for transfer learning was released by the HIT-SCIR and iFLYTEK joint laboratory. This model was trained on Chinese Wikipedia (https://dumps.wikimedia.org/zhwiki/latest/) and other general corpora (BQ corpus, ChnSentiCorp, CJRC, CMRC2018, LCQMC, MSRA, PFR, XNLI). Notably, Wikipedia is edited by global knowledge contributors, ensuring data quality. However, as the internet is a relatively open platform, data generated on it often contains considerable noise that would directly impact model accuracy. Therefore, the model strictly filtered supplementary corpora during pretraining—all supplementary corpora are publicly available datasets widely used in the NLP field, carefully compiled by researchers with high quality, thus largely avoiding negative impacts from low-quality data.

Experimental hardware specifications: GPU: Quadro RTX 8000; VRAM: 48G; CPU: 2×Xeon Platinum 8160; RAM: 128G. Software environment: OS: Ubuntu 16.04 LTS, NVIDIA driver: 418.56, CUDA version: 10.1, programming language: Python 3.7, deep learning framework: PyTorch 1.5.

### 4.2 Experimental Design

To observe the actual effectiveness of our constructed model for domain information identification, we designed three control experiments: (1) using a RoBERTa model pretrained only on the training set (i.e., without transfer learning) combined with DPCNN for domain information identification; (2) using only the transfer-learned RoBERTa for domain information identification; (3) using the classic deep learning model TextCNN. The final experimental framework is shown in Figure 7 [Figure 7: see original paper].

Precision, recall, and F1-score are adopted as evaluation metrics. Precision refers to the proportion of true positive samples among those predicted as positive by the model. Recall refers to the proportion of true positive samples

correctly identified by the model. F1-score is the harmonic mean of the two. The three metrics are calculated as follows:

$$\text{Precision } P = TP/(TP + FP)$$
$$\text{Recall } R = TP/(TP + FN)$$
$$F1 = 2PR/(P + R)$$

In formulas (3) and (4), TP represents the number of true positive samples predicted as positive, FP represents the number of true negative samples predicted as positive, and FN represents the number of true positive samples predicted as negative.

### 4.3 Experimental Results Analysis

Our constructed model and baseline models were used to identify and evaluate domain information on the test set. The precision, recall, and F1-scores of each model were calculated, statistically analyzed, and summarized in Table 1 . In the table, "TL" and "NoTL" represent RoBERTa models with and without transfer learning, respectively, and "D" denotes the DPCNN model.

**Table 1: Model Performance Evaluation**

[Table content showing precision, recall, and F1 scores for TL+D, NoTL+D, and TextCNN across 10 domains]

Comparing the experimental group <TL+D, NoTL+D, TextCNN>, we can clearly observe that the RoBERTa+DPCNN model without transfer learning performs slightly worse than the classic deep learning model TextCNN. However, after applying transfer learning, all metrics improve significantly. When the highest values under each statistical indicator across domains in Table 1 are bolded, the transfer learning-based RoBERTa+DPCNN model almost exclusively occupies the top positions. Compared with the classic TextCNN model, the average absolute improvements across domains are 3.43% in precision, 3.44% in recall, and 3.44% in F1-score. Compared with the non-transfer learning RoBERTa+DPCNN model, the improvements are 4.15%, 4.55%, and 4.52%, respectively. These data fully demonstrate that introducing transfer learning in domain information identification can fully leverage big data advantages and improve model identification performance.

Notably, when identifying education domain information, the RoBERTa+DPCNN model trained only on the training set achieves slightly higher precision than the transfer learning model. However, observing the recall metric reveals that this slight precision improvement comes at the cost of substantially sacrificing recall, which is difficult to accept for domain information identification tasks as it would miss considerable domain-relevant information. Consequently, its F1-score is lower than that of the transfer learning model. Additionally,

the "NoTL+D" model shows generally low recall across all data. Through further analysis, we believe this model, with its large number of parameters, is a complex model, while the training set sample size is small. When the problem being addressed is also complex, the model's parameters are in an "under-learning" state, insufficient for fitting enough features to make good domain information judgments. However, in a transfer learning environment, this problem does not exist due to sufficient corpora supporting model training.

Continuing to observe the <TL+D, TL, TextCNN> experimental group, we find that the transfer learning model surpasses the traditional deep learning model TextCNN in almost all metrics. Using only the pretrained RoBERTa model achieves average improvements of 3.34%, 2.56%, and 2.46% over TextCNN in precision, recall, and F1-score, respectively. Observing only the <TL+D, TL> group, we find that processing transfer learning results with DPCNN further improves model performance, with average improvements of 1.09%, 0.88%, and 0.98% in precision, recall, and F1-score, respectively. This indicates that further refined adjustment of pretrained model outputs can enhance model performance and avoid wasting transfer learning results.

Furthermore, further manual analysis of the identification results reveals: (1) Compared with non-transfer learning models, transfer learning models demonstrate better identification performance for information without obvious domain characteristics in the text. For example, a current affairs news article stating "Our country will continue to increase punishment for pyramid scheme crimes" contains no explicit policy or situation-related keywords, but semantically belongs to the current affairs domain. Only the <TL+D, TL> transfer learning models correctly identified this information. (2) For domain information that may contain multiple themes, all models show poorer identification results. For example, the news article "Revealing Popular Majors: Employment Reality of Finance and Foreign Trade Majors" actually analyzes popular majors and should belong to the education domain, although the major is related to finance. Consequently, all models failed to identify this news when identifying education information, while incorrectly identifying it when identifying finance information.

For domain information identification, real-world application environments are more complex, with the amount of information covered in target domains being far lower than the total information in the real space—that is, the data volumes are highly imbalanced. To further evaluate model accuracy and generalization ability under imbalanced sample spaces, we plotted ROC (Receiver Operating Characteristic) curves for each model across the ten domains, as shown in Figure 8 [Figure 8: see original paper].

The value in parentheses after each model in the figure represents the AUC (Area Under Curve) covered by the ROC curve. This value serves as both a reference for model accuracy assessment and largely represents model performance on imbalanced datasets. The figure shows that all tested models achieve relatively high AUC values, but overall, transfer learning models demon-

strate better generalization ability on imbalanced datasets. After averaging each model's AUC across all domains, the values for TL-RoBERTa+DPCNN, NoTL-RoBERTa+DPCNN, TL-RoBERTa, and TextCNN are 0.990, 0.981, 0.985, and 0.986, respectively, again indicating that our constructed model performs best and possesses superior generalization capability.

## Conclusion

Addressing the problem that domain information identification performance is difficult to improve due to limited learning samples, this paper proposes a method using transfer learning for more accurate domain information identification. Experimental results demonstrate that the transfer learning-based RoBERTa+DPCNN model substantially outperforms both non-transfer learning models and the classic TextCNN deep learning model, with average absolute improvements of 4.15% and 3.43% in precision, 4.55% and 3.44% in recall, and 4.52% and 3.44% in F1-score, fully proving the effectiveness of transfer learning. Additionally, aggregating and condensing the transfer learning model's output with DPCNN improves precision, recall, and F1-score by 1.09%, 0.88%, and 0.98% on average, respectively, indicating that refined adjustment of transfer learning results enhances model identification performance.

The paper also has limitations. When fine-tuning the transfer learning model, we retained all parameters of the original pretrained model. However, research [25] has shown that selective parameter adjustment of transfer models can yield better performance in practical tasks. In future work, we will conduct more specific research on this issue to achieve even better domain information identification results.

## References

[1] Ringel D, Radinsky K, Markovitch S. Cross-cultural transfer learning for text classification[D]. Israel: Technion, 2019.
[2] Yu S, Su J, Luo D. Improving BERT-based text classification with auxiliary sentence and domain knowledge[J]. IEEE access, 2019, 7: 176600-176612.
[3] Pan Hongliang, Wang Zhengde. Information Knowledge Dictionary[M]. Beijing: Military Friendship Press, 2002.
[4] Zhang Xuegong. Pattern Recognition[M]. 3rd ed. Beijing: Tsinghua University Press, 2010.
[5] Ma Y, Tang J, Aggarwal C. Feature engineering for data streams[M]//Feature engineering for machine learning and data analytics. Boca Raton: CRC Press, 2018: 117-143.
[6] Liao Liefa, Le Fugang, Zhu Yalan. Application of LDA model in patent text classification[J]. Modern Information, 2017, 37(3): 35-39.
[7] Yang Tengfei, Xie Jibo, Li Zhenyu, et al. Identification and classification of typhoon disaster loss information in Weibo[J]. Journal of Geo-Information Science, 2018, 20(7): 906-917.

[8] Kim Y. Convolutional neural networks for sentence classification[C]//Yuval M. Empirical methods in natural language processing. Qatar: ACL, 2014: 1746-1751.
[9] Huang Tao. Research and implementation of news classification system based on machine learning[D]. Beijing: Beijing University of Posts and Telecommunications, 2019.
[10] Liu P, Qiu X, Huang X. Recurrent neural network for text classification with multi-task learning[C]//IJCAI. Proceedings of the twenty-fifth international joint conference on artificial intelligence. New York: AAAI Press, 2016: 2873-2879.
[11] Joulin A, Grave E, Bojanowski P, et al. Bag of tricks for efficient text classification[C]//Mirella L. the 15th conference of the european chapter of the association for computational linguistics. Spain: EACL, 2017: 427-431.
[12] Johnson R, Zhang T. Deep pyramid convolutional neural networks for text categorization[C]//Hinrichs S. Proceedings of the 55th annual meeting of the association for computational linguistics. Vancouver: ACL, 2017: 562-570.
[13] Zhuang Fuzhen, Luo Ping, He Qing, et al. Research progress on transfer learning[J]. Journal of Software, 2015, 26(1): 26-39.
[14] Zhou B, Lapedriza A, Xiao J, et al. Learning deep features for scene recognition using places database[C]//Roman G. In advances in neural information processing systems. Harrahs: NIPS, 2013: 3111-3119.
[15] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Isabelle G. Advances in neural information processing systems. Cambridge: MIT Press, 2014: 487-495.
[16] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Isabelle G. Advances in neural information processing systems. Long Beach: NIPS, 2017: 5998-6008.
[17] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.
[18] Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J/OL]. [2020-06-28]. https://arxiv.org/pdf/1810.04805.
[19] Liu Y, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach[J/OL]. [2020-04-01]. https://arxiv.org/pdf/1907.11692.
[20] Lan Z, Chen M, Goodman S, et al. ALBERT: a lite BERT for self-supervised learning of language representations[J/OL]. [2020-04-01]. https://arxiv.org/pdf/1909.11942.
[21] Houlsby N, Giurgiu A, Jastrzebski S, et al. Parameter-efficient transfer learning for NLP[J/OL]. [2020-04-01]. https://arxiv.org/pdf/1902.00751.
[22] Sharma T, Upadhyay U, Bagler G. Classification of cuisines from sequentially structured recipes[C]//2020 IEEE 36th international conference on data engineering workshops (ICDEW). Dallas: IEEE, 2020: 105-108.
[23] Sun Maosong, Li Jingyang, Guo Zhipan, et al. THUCTC: an efficient Chinese text classification toolkit[EB/OL]. [2020-05-10]. http://thuctc.thunlp.org/.
[24] Cui Y, Che W, Liu T, et al. Revisiting pre-trained models for Chinese natural language processing[J/OL]. [2020-06-01]. https://arxiv.org/pdf/2004.13922.

[25] Tenney I, Das D, Pavlick E. BERT rediscovers the classical NLP pipeline[J/OL]. [2020-01-01]. https://arxiv.org/pdf/1905.05950.

## Author Contributions

Lu Quan: Proposed research ideas and designed paper framework
Hao Zhitong: Model construction and experiments, paper writing
Chen Jing: Designed research methodology, revised paper
Chen Shi: Data preprocessing
Zhu Anqi: Experimental design

---

**Discussion on Using Transfer Learning to Accurately Identify Domain Information**

Lu Quan[1],[2], Hao Zhitong[1], Chen Jing[3], Chen Shi[1], Zhu Anqi[1]

[1]Center for Studies of Information Resources, Wuhan University, Wuhan 430072
[2]Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources, Shenzhen 518034
[3]School of Information Management, Central China Normal University, Wuhan 430079

**Abstract:** [Purpose/significance] To solve the problem that the identification effect of target domain information is difficult to improve due to insufficient learning samples, we transfer the results of unsupervised learning from big data to the feature space of the target domain. [Method/process] We used the RoBERTa model, which was pretrained with Chinese Wikipedia and other data, for transfer learning. After mapping the learning results to the target domain, DPCNN was used to aggregate and condense it, and then the model was fine-tuned with part of the labeled data to complete the accurate recognition of domain information. [Result/conclusion] Compared with the model without transfer learning and the classic model TextCNN in 10 fields, the model in this paper is much better than the comparison models. After average, the precision is increased by 4.15% and 3.43%, the recall is increased by 4.55% and 3.44%, and the F1 score is increased by 4.52% and 3.44%. It shows that knowledge transfer using big data can effectively improve the information recognition effect in the target field.

**Keywords:** transfer learning; information recognition; RoBERTa

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