A Few-Shot Transfer Learning Approach Using Text-label Embedding with Legal Attributes for Law Article Prediction

Yuh-Shyan Chen, Shin-Wei Chiang and Tong-Ying Juang

July 30, 2019

A Few-Shot Transfer Learning Approach Using Text-label Embedding with Legal Attributes for Law Article Prediction

Yuh-Shyan Chen, Shin-Wei Chiang and Tong-Ying Juang
Department of Computer Science and Information Engineering, National Taipei University
Sanxia, Taiwan, 237, R.O.C.
E-mail: yschen@cs.ntpu.edu.tw

***Abstract*— Law article prediction is to determine the appropriate law article according to the fact descriptions, which is useful for legal auxiliary system when given a fact description of the case. A problem of legal charge prediction is often faced with the imbalance data problem, which represent the few-shot category with limited case. To address the imbalance data, this paper introduces law article as label embedding to improve the relevance between fact and law articles. More specifically, this paper applies the weight sharing mechanism of transfer learning to utilize the data with high frequency to model the projection between fact and law articles, also as a prior knowledge to achieve law article classification for case with low frequency data. This paper employs an attention mechanism based on text similarity to produce the fact context vector, and then infer the law article associated with the case by utilize a label set independent projection layer. The experimental results show that, the label embedding of law article can improve the prediction performance, and transfer the weight of projection layer to few-shot data can achieves better performance on few-shot data classification.**

Index Terms: Natural language processing, deep learning, few-shot learning, law article prediction, legal judgment prediction.

I. Introduction

In the judicial field, the court's verdict is related to whether the case cited the law is correct, so lawyers and prosecutors must carefully choose the appropriate law. However, how to quote the correct law for each case is not a simple matter and requires a lot of experience and the time cost of querying the information. In recent years, the development of neural networks has enabled us to use the text of the case to predict the law that the case should be cited and to provide an effective solution to the legal judgment.

In recent years, neural networks have achieved great success in the field of natural language processing (NLP) . Yoon *et al.* [1] propose to use convolutional neural networks for classification which convert text to a fixed-dimensional vector and extract features through the convolution kernel. Chen *et al.* [2] is recommended to represent the semantic information of the text and model the high-order label correlation by combining CNN with RNN. However, these method above, either do not take into account the relationship between labels or do not consider differences in the importance of textual content when predicting labels.

The recent success of deep learning heavily rely on a large number of labeled training data. However, data collection is accompanied by a large amount of labor costs, and in real case, there are often faced with category imbalances, for example, unusual diseases in the medical field or uncommon cases in the legal field. Supervised learning algorithm are more difficult to achieve good performance in this situation.

In view of above situation, transfer learning [3] and zero-shot learning [4] are considered to be a promising solution. Transfer learning is a machine learning method where a model developed for a task is reused as the base point for a model on a second task. Improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned. Zero-shot learning is a particular form of transfer learning. This is used to solve the learning task with rare sample and unbalanced situation, to utilize the source domain with a large labeled data to learn the feature relation and transfer to the task with only the few-shot sample data.

This paper combines the above research and applies it to the legal field. Wisdom law is a new application of deep learning and law. Existing works generally regard wisdom law as a text classification problem [5]. The goal of smart law is to establish a legal aid system based on the facts of the case to predict the outcome of the judgment (e.g., relevant law articles, charges, prison terms, etc.) .

Law article prediction is the main goal of this work. Due to the problem of imbalance data in the law article dataset, this leads to low frequency categories that are difficult to predict. In order to solve the imbalance data problem, this paper uses the weighting mechanism based on text similarity to generate the factual context vector, and uses the weight sharing approach of transfer learning as the prior knowledge of few-shot data training, and proposes a classification layer that is not limited by the number of categories.

The rest of this paper is organized as follows. Section II introduces the related researches, section III describes the basic ideal and problem formulation of our schema. The proposed schema presented in section IV. Section V gives the performance analysis. Finally, section VI concludes this paper.

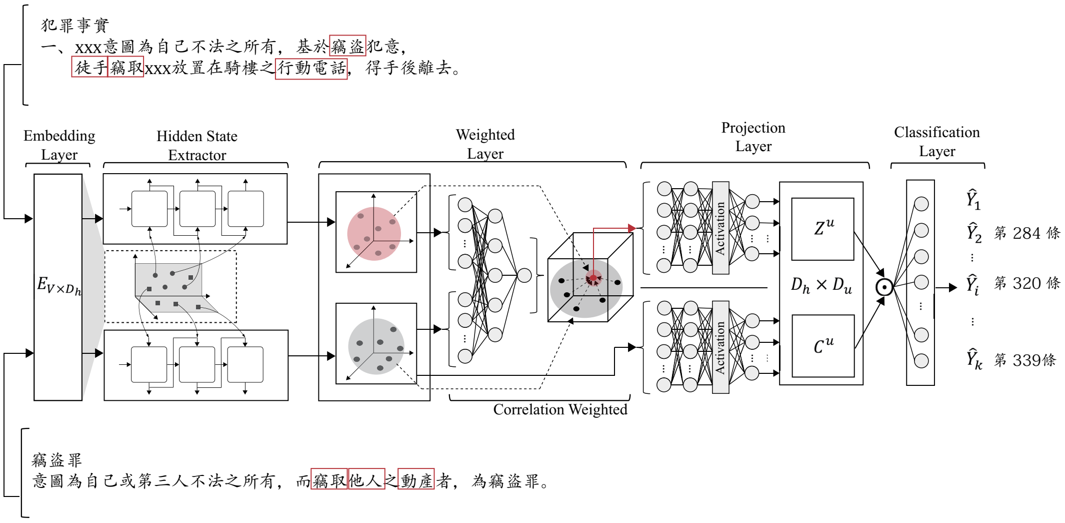


Fig. 1: The law article prediction approach architecture.

II. Related Works

The few-shot learning is a kind of application in transfer learning, most of the few-shot researches are focused on computer visual aspects. The common method of few-shot task is to utilize the attributes of the category, that can offer an intermediate representation among different categories. Lampert *et al.* [6] proposes an attribute classifier which integrate direct attribute prediction(DAP) and indirect attribute prediction (IAP) to finding the new suitable object class and dont need re-training. In the field of text classification, Su *et al.* [7] has proposed a approach for open text classification which based on deep learning, the goal of the approach is to recognize new documents during test phase but the approach is not capable of unseen classes.

In the legal intelligence field, legal judgment prediction is a hot topic which has drawn attention for decades. In early years, the researches of legal judgment prediction (LJP) was based on the statistical and mathematical methods, recent researches are employing deep learning techniques to legal judgment prediction (LJP) and has made good progress [8][9].

In former days, with the development of deep learning and natural language processing technology, more researches tend to treat the task as a text classification problem and absorbed in two themes, judgment prediction and charge prediction.

The goal of judgment prediction is to predict the court decision based on the fact of the case. Zhong *et al.* [10] proposes a judgment prediction framework named TOPJUDGE which use the topological multi-task learning framework to explore and formalize the multiple sub tasks of legal judgment. Long *et al.* [11] treats the judgment prediction task to the reading comprehension problem, the proposed schema can models the better complementary input.

Charge prediction aims to utilize the fact description of the case to predict the final charge. Hu *et al.* [12] employs an attribute attention mechanism to establish an interaction between attribute and fact description, also proposed a multi-task learning framework to predict the charge of the case. Jiang *et al.* [13] introduces a reinforcement learning method to extract rationale of the fact, also solves the interpret performance problem existing in charge. Ye *et al.* [14] proposes a attention based sequence-to-sequence model to generate court views used charge labels, and releases the data set used in proposed work. Luo *et al.* [15] proposes a multi-task framework to extract the relevant article of the case and utilize then to model the charge prediction task jointly. However, the researches mostly of focuses on high-frequency case, without paying attention to rare cases, therefore, this paper proposes a weight sharing projection network schema with attention mechanism to address these issues.

III. Preliminaries

This section describes the system model, the problem formulation, and basic idea in subsections III.A, III.B and III.C.

A. System model

The proposed schema in this paper reaches the few shot prediction through four steps, first step is to construct a trainable non-static word embedding matrix and learning more suitable vector representation through training. To consider the importance of different words in the text, the attention mechanism based on vector similarity approach is used in step two. For the few shot learning, step three employs a projection network to make the weight update of the classification layer is not restricted by the category. Finally, a classification layer that shares weights between different categories is used. The model structure is shown in Fig. 1.

B. Problem Formulation

The main goal of our proposed model is given a fact description X_i and then predict the charge Y_i associated with the fact X_i. It can be regards as a text classification problem with single or multi-labels, based on this idea, the training objective is to minimize the classification loss, it can be formalized as:

$$min\frac{1}{N}\sum_{n=1}^{N}H(Y_n,\hat{Y_n})\ \ , \tag{1}$$

where $H\left(Y_n,\hat{Y_n}\right)$ is the cross-entropy between the Ground-truth label distribution Y and predicted label probability distribution $\hat{Y_n}$. The formula can be expand as:

$$\begin{gathered} H\left(Y_n,\hat{Y_n}\right)= \\ -\frac{1}{m}\sum_{j=1}^{m}y_{n,j}\cdot log\hat{y}_{n,j}+(1-y_{n,j})\cdot log(1-\hat{y}_{n,j}) \\ subject\ to\begin{cases} y_{n,j}\in\{0,1\}\,, & \forall n,j \\ 0\leq\hat{y}_{n,j}\leq 1, & \forall n,j \end{cases} \end{gathered} \tag{2}$$

where $y_{n,j}$ is denote *j*-th label in the ground-true label set Y_n, and $\hat{y}_{n,j}$ denote predicted label as the same representation.

In the evaluation stage, our expectation is to pursuit the maximize model accuracy, the accuracy is calculated by predicting the correct rate of the label, the formula as shown in Eq. 3 :

$$\begin{gathered} max\frac{1}{N}\sum_{n=1}^{N}\frac{\sum_{k=1}^{m}(y_{n,k}\times\hat{y}_{n,k})}{m}, \\ subject\ to\begin{cases} 0\leq\hat{y}_{n,k}\leq 1, & \forall n,k \\ if\quad \hat{y}_{n,k}>0.5, & 1 \\ else\ \ \hat{y}_{n,k}\leq 0.5, & 0 \end{cases} \end{gathered} \tag{3}$$

In the formula, when the predicted label probability is higher then the threshold, the label is treated as a prediction label, otherwise treated as unrelated label.

C. Basic ideal

The basic idea of the proposed model is manly to handle the few-shot text classification problem utilize label attributes. The traditional classification model done this thing used a fully connected network as a classification layer which the network parameters to be related to the number of target label. When the training data set have unbalanced distribution, the weight parameters corresponding to a small number of categories will rarely be updated and also unable to face the unseen label.

IV. A Few-Shot Transfer Learning Approach

This section presents an using projection network with transfer weight sharing as prior knowledge to predict few shot category in the filed of legal. An overview of the proposed scheme is given, there are four phases, presentation as follow:

A. Document vectorization phase

The data set used in this paper is the indictment and judgment of the Taiwan criminal law collected from the government open platform. Each document contains a fact description and the related law article of the case. Because it is a real-world data set, the content contains a lot of noise, and needs to be cleaned up and statistically analyzed. According to the number of occurrences of the law, the data set is divided into high-frequency law data sets and low-frequency law data sets, which are used for the follow model training and verification. The processing of data corpus includes data cleaning, word segmentation, and conversion of text into dictionary index according to word frequency statistics, and then establish the embedded matrix and training word vector. The steps are detailed below:

S1: Since there is no similar data set in the field of legal research in Taiwan, the data set of this paper is collected from the government open platform. Our data set consists of criminal justice cases in Taiwan. Each data consists of two fields, fact description X and the corresponding reference related law Y. This study focuses on the prediction of laws related to fact descriptions, filter out the laws related to the judicial process, according to the frequency of occurrence of the law, the data set is divided into high frequency data sets and low frequency data sets.

S2: Segment the words in the fact description X_i of data set $D=(X_i,Y_i)$ and the text description of law article $\{A_j|j=1...k...m\}$, the word segmentation framework is used in this paper is HanLP. Establishing a dictionary $\{word: The\ number\ of\ occurrences\}$ according to the word frequency, then allows to vectorize a text corpus, by turning each text into either a sequence of integers.

S3: Previous steps convert a fact description into a text sequence which composed of index of word tokens $X_i=\{w_{1,1},w_{1,2},...w_{i,j},...w_{s,t}\}$. In order to generate a word embedding vector, this paper introduce a pre-train embedding model which is trained by the Word2Vec [16]. The $embedding\ matrix\in R^{V\times D}$ is constructed by select the word of top n frequently occurring words in the dictionary mentioned before, n is selected depending on the task. For each sample, the text sequence $X_i=\{w_{1,1},w_{1,2},...w_{i,j},...w_{s,t}\}$ have been embedded to the vector representation $V_i=\{v_{1,1},v_{1,2},...v_{i,j},...v_{s,t}\}$, where $v_j\in R^D$, D is the dimensionality of the embedding space.

B. Context correlation weighting phase

How to produce a meaningful text representation of vector is the key in text classification. This phase joint embedded the words and labels to the vector representation.To give different influence according to the relationship between word and labels, and to produce a meaningful context vector z. Different from the traditional attention mechanism to learn the weight of the token through training, we hope to learn more suitable embedding vectors directly. The details as Fig. 2.

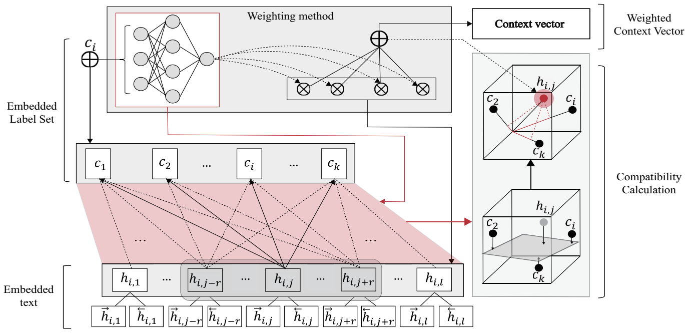


Fig. 2: The context correlation weighting approcah

S1: In this step, first thing is to encode the attribute of the law $A_j = \{a_{j,k}\}_{k=1}^{m}$ to the context vector of the fixed dimension $c_j \in R^D$ as label vector, where D is the dimension of vector, the encode function formulation as Eq. 4, where l_j is the sequence length of law attribute A_j and m is the number types of law, here, considers that the length of the laws attribute are short, so only the vector average is used for encode. The encoded label vector is arranged in a matrix, such as Eq. 5.

$$c_j = \frac{1}{l_j}\sum_{k=1}^{l_j} a_{j,k},\ \forall j = 1...m\ \ , \tag{4}$$

$$C = \begin{bmatrix} [c_{1,1},\ c_{1,2}\ ...c_{1,k},\ ...c_{1,d}] \\ \vdots \\ [c_{j,1},\ c_{j,2}\ ...c_{j,k},\ ...c_{j,d}] \\ \vdots \\ [c_{m,1}, c_{m,2}...c_{m,k}, ...c_{m,d}] \end{bmatrix},\ C \in R^{m\times D} \tag{5}$$

S2: In order to encode the label associated input text, this phase focus on the Bi-LSTM network, the network take as input a vectorized text sequence V and output a weighted context output z. The weighted method is based on vector similarity, We extract the hidden vector of each time step both of forward LSTM $\overrightarrow{h}$ and backward LSTM $\overleftarrow{h}$ to generate the hidden vector $h_{i,j}$, which is represent the $j-th$ word context vector, as shown in Eq. 6. Where $\overrightarrow{h}$, $\overleftarrow{h} \in R^D$ and $h_{i,j} \in R^{2D}$.

$$\begin{aligned} \overrightarrow{h}_{i,j} &= \overrightarrow{LSTM}\left(\overrightarrow{h}_{i,j-1}, v_{i,j}\right) \\ \overleftarrow{h}_{i,l-j} &= \overleftarrow{LSTM}\left(\overleftarrow{h}_{i,l-j+1}, v_{i,l-j}\right) \\ h_{i,j} &= \overrightarrow{h}_{i,j} \oplus \overleftarrow{h}_{i,j} \end{aligned} \tag{6}$$

The V has been encoder to $H_i = \{h_{i,j}\}_{j=1}^{l}$, the similarity weighted compute as below, for each word in H, to compute the similarity with the label matrix which obtained in S1, the formula of similarity measure is shown Eq. 7, where c_k is the k-th element in label matrix C.

$$Cosine\,(h, c_k) = \frac{h \cdot c_k}{\|h\| \times \|c_k\|},\ k = 1...m \tag{7}$$

The single $s_{j,k}$ defined as the similarity between word h_j and label c_k, therefore, after calculating by Eq. 7, the similarity matrix S can be obtained, as shown in Eq. 8, which dimension of the matrix is $m \times l$, where l is the length of text, m is the number of label.

$$S = \begin{bmatrix} s_{1,1} & s_{2,1} & & & \\ s_{1,2} & \ddots & & \ddots & \\ \vdots & & s_{j,k} & & \vdots \\ & \ddots & & \ddots & s_{l,m-1} \\ & & & s_{l-1,m} & s_{l,m} \end{bmatrix} \tag{8}$$

S3: In order to take into account the semantics between adjacent words (phrase level) and capture the interaction between words, here, one-dimensional convolution calculation is performed on the similarity matrix obtained by S2. The size of the convolution kernel β can be regarded as the length of the phrase, and the interaction relationship between each word is considered at the same time. The output is a similarity representation of the phrase to the label, where j is a phrase start from j-th words. Formula such as Eq. 9.

$$p_j = \sigma\left(\theta \cdot S_{j:j+\beta-1} + b\right) \tag{9}$$

A convolution operation involve a filter $\theta \in R^{\beta\times m}$, $b \in R$ is a trainable bias term, σ is a non-linear function and $p_j \in R^m$. The input phase-level similarity matrix $S = \{s_{1:1+\beta}, s_{2:\beta+1}...s_{l-\beta+1:l}\}$ have been produced the feature map $P = \{p_1, p_2...p_l\}$ by the convolution filter. Then employ a max-pooling operation over p_j to extract the largest value of the phrase-level similarity representation wrt the labels, the operation is Eq. 10.

$$m_j = MaxPooling(p_j) \tag{10}$$

As a result of Eq. 10, $M = [m_1, m_2..., m_l]$ as the importance for each word in text. To calculate the attention score for input text, here first apply $softmax$ operation (Eq. 11) over M then generate the weighted context vector by Eq. 13.

$$\alpha_j = \frac{exp\,(m_j)}{\sum_{t=1}^{l} exp\,(m_t)} \tag{11}$$

$$z = \sum_{j=1}^{l} \alpha_j v_j \tag{12}$$

Finally, $z \in R^D$ is the weighted context vector which is implied the label interaction.

C. Fact-label projection phase

The phase II obtains the weighted vector $z \in R^d$. In order to enable the model to extend the prediction of rare category, a projection layer is used in phase III, the weight of the classification layer is independent of the number of

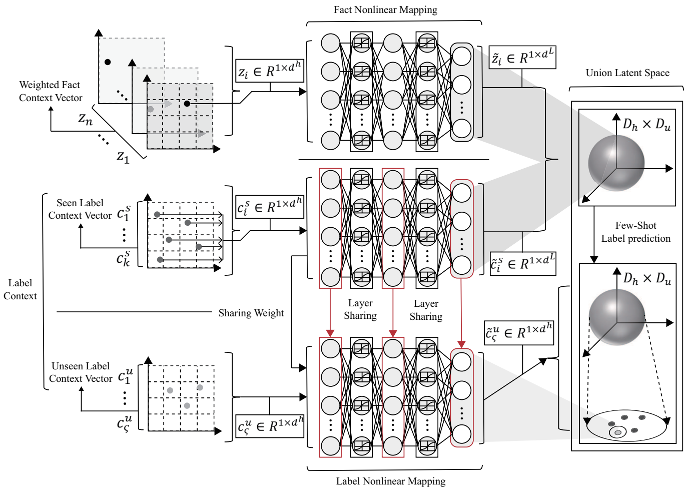


Fig. 3: Fact-label projection phase

categories. The weight of the model learns the interactions in the vector space, rather than the classification relationships for the categories, that allowing the model to achieve a rare number of categories of predictions, the detail as shown in Fig. 3.

S1. The general text classification model is structured to generate a context vector of a text message and then classify the category by a nonlinear model. This way, the classification weight is limited by the number of categories. Taking the legal data set as an example, the weight of the rare law will be rarely updated. Therefore, this paper uses a classification layer that is independent of the number of categories. The implementation method is to pass the weighted context vector z generated by phase II and the label matrix C through two embedded networks $E_{text}(z) = z^u$ and $E_{label}(c) = c^u$ to mapping the text vector z and label vector c into a text-label union space $U \in R^{d \times d_u}$ respectively, the embedding formula as Eq. 13 and Eq. 14.

$$\begin{aligned} E_{text}(z) = max\left(0, z \cdot w_{text} + b_{text}\right) \quad , \\ subject\ to \begin{cases} 0, & z \cdot w_{text} + b_{text} \leq 0 \\ 1, & z \cdot w_{text} + b_{text} > 0 \end{cases} \end{aligned} \tag{13}$$

$$\begin{aligned} E_{label}(c) = max\left(0, c \cdot w_{label} + b_{label}\right) \quad , \\ subject\ to \begin{cases} 0, & c \cdot w_{label} + b_{label} \leq 0 \\ 1, & c \cdot w_{label} + b_{label} > 0 \end{cases} \end{aligned} \tag{14}$$

After encoding each label vector $c_j | j = 1,,,m$, the encoded label vector matrix C^u can be obtained, such as Eq. 15.

$$C^u = \begin{bmatrix} E_{label}(c_1) = c_1^u \\ \vdots \\ E_{label}(c_j) = c_j^u \\ \vdots \\ E_{label}(c_m) = c_m^u \end{bmatrix} \tag{15}$$

S2. Let the weight of the classification layer learn the interaction between text and label in union space U, and establish the structure of space. Here we use the hadamard product of text z^u and label c^u to express the relationship between the two, as Eq. 16.

$$g_j^u = \left\{ \left(z^u\right)^\top \cdot c_j^u \right\}_{j=1}^{m} \quad , \tag{16}$$

where $g_j^u \in R^{d_u}$ is denote the compatibility between encode text $z^u \in R^d$ and any known encode label $c_j^u \in R^d \mid j = 1..., k, ..., m$, m is the number of known label. Extended to matrix operation, the compatibility between z^u and encode label matrix $C^u \in R^{m \times d}$ can be express as Eq. 17.

$$G^u = \begin{bmatrix} g_1^u \\ \vdots \\ g_j^u \\ \vdots \\ g_m^u \end{bmatrix}^{m \times D_u} = \begin{bmatrix} \left(z^u\right)^\top \cdot c_1^u \\ \vdots \\ \left(z^u\right)^\top \cdot c_j^u \\ \vdots \\ \left(z^u\right)^\top \cdot c_m^u \end{bmatrix} \tag{17}$$

S3. The classifier is implemented as multi-layer fully connected layers (FC) , the FC output a score of text z belong to k-th known label which according to the compatibility g_j^u calculated from s2, and model the linear mapping in the union space, the FC did not use non-linear activation. The weight of the FC learns to model the structure of the union space according to the correlation g which computed by Eq. 16, rather than the type of the label, therefore, the classification layer irrelevant to the number of label can be achieved. The score calculation can be formula as Eq. 21. For each label shares the same set of mapping weights, prediction for all known labels which is belong to text z_j are as Eq. 19:

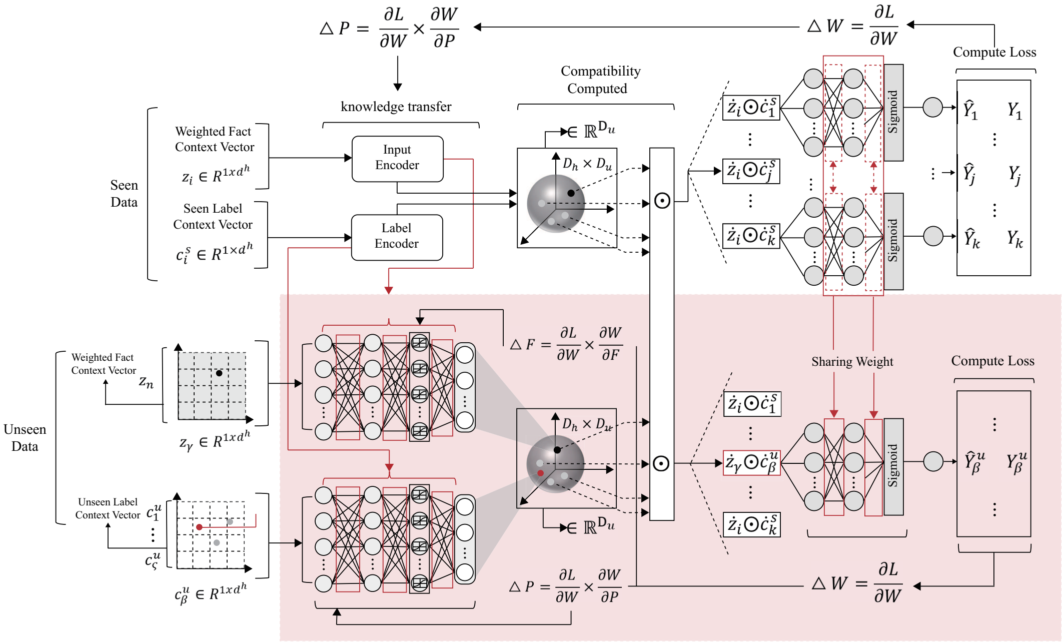
$$score_{i,j} = g_{i,j}^u \cdot w_c + b \quad , \tag{18}$$

$$Score_i = \begin{bmatrix} score_{i,1} \\ \vdots \\ score_{i,j} \\ \vdots \\ score_{i,m} \end{bmatrix} = \begin{bmatrix} g_{i,1}^u \cdot w_c + b \\ \vdots \\ g_{i,j}^u \cdot w_c + b \\ \vdots \\ g_{i,m}^u \cdot w_c + b \end{bmatrix} \quad , \tag{19}$$

To convert the score to the valid probability estimate and calculated independently of each other, this paper apply a sigmoid function to compression the prediction probability from 0 to 1, the formula as fellow :

$$\begin{aligned} \hat{Y}_i = P(Y_i | X_i) \\ P(Y_i | X_i) = \frac{1}{1 + e^{-(Score_i)}} \quad , \end{aligned} \tag{20}$$

where $\hat{Y}_i = \{\hat{y}_{i,j}\}_{j=1}^{m}$ is the prediction probability of the model, which is a m-dimension vector, each instance



$\triangle P = \frac{\partial L}{\partial W} \times \frac{\partial W}{\partial P}$
$\triangle W = \frac{\partial L}{\partial W}$
knowledge transfer
Compatibility Computed
Compute Loss
Seen Data
Weighted Fact Context Vector
$z_i \in R^{1xd^h}$
Seen Label Context Vector
$c_i^s \in R^{1\times d^h}$
Input Encoder
Label Encoder
$\in \mathbb{R}^{D_u}$
$D_h \times D_u$
$\dot{z}_i \odot \dot{c}_1^s$
$\dot{z}_i \odot \dot{c}_j^s$
$\dot{z}_i \odot \dot{c}_k^s$
Sigmoid
$\hat{Y}_1$ Y_1
$\hat{Y}_j$ Y_j
$\hat{Y}_k$ Y_k
$\triangle F = \frac{\partial L}{\partial W} \times \frac{\partial W}{\partial F}$
Sharing Weight
Compute Loss
Unseen Data
Weighted Fact Context Vector
z_n
$z_\gamma \in R^{1xd^h}$
Unseen Label Context Vector
c_1^u
c_ς^u
$c_\beta^u \in R^{1xd^h}$
$\dot{z}_\gamma \odot \dot{c}_\beta^u$
$\hat{Y}_\beta^u$ Y_β^u
$\triangle P = \frac{\partial L}{\partial W} \times \frac{\partial W}{\partial P}$
$\triangle W = \frac{\partial L}{\partial W}$

Fig. 4: Few-shot classification by transfer learning phases.

of the vector $\hat{Y}_i$ represent the probability of the text X_i belong to that label, when in evaluation step, if $\hat{Y}_i$ greater than the threshold then determination the label is assigned. To summary the previous operation.

D. *Few-shot classification by transfer learning phases*

The model proposed in this paper apply the concept of domain transfer from transfer learning. The purpose is to use the larger number of category data to establish the mapping relationship of the classification layer as external knowledge, share the mapping relationship in the embedded space and migrate to a small number of categories of data. The objects classified by the classification layer are the correlations between vectors. After the calculation of phase3, the weight of the neural network is shared for each category in the classification layer, so as to achieve the purpose of migrating external knowledge to a small number of categories, as shown in Fig. 4.

S1. To use the data set L_s with a large number of categories to train, establish the context vector and embedding space associated with the label through phaseII, and establish the mapping relationship of union space through phaseIII. Finally, a FC classifier that classifies tasks according to the mapping relationship between categories is trained. The FC classifier shares weights among different categories, so the network structure of learning can be restricted by the number of categories. Training objectives can be set as:

$$\begin{aligned}\pounds\left(\Theta\right) = & \frac{1}{N}\sum_{n=1}^{N} H\left(Y_n, \hat{Y}_n\right) + \\ & \lambda\frac{1}{m}\sum_{j=1}^{m} H\left(y_j, F_2 \circ F_1\left(c_j, E_{text}\left(c_j\right)\right)\right)\end{aligned} \tag{21}$$

$$\begin{aligned} & H\left(Y_n, \hat{Y}_n\right) = \\ & -\frac{1}{m}\sum_{j=1}^{m} y_{n,j} \cdot log\hat{y}_{n,j} + (1 - y_{n,j}) \cdot log(1 - \hat{y}_{n,j}) \\ & subject\ to \begin{cases} y_{n,j} \in \{0,1\}, & \forall n, j \\ 0 \le \hat{y}_{n,j} \le 1, & \forall n, j \end{cases} \end{aligned} \tag{22}$$

where $H\left(Y_n, \hat{Y}_n\right)$ is the cross-entropy function formula as Eq. 22. Y_n is the actual label of case n and $\hat{Y}_n$ is the prediction label probability. In order to make the shared classification layer weights not over-fitting when migrating to different categories, restrictions are added to the formula, such as Eq. 23.

$$\lambda\frac{1}{m}\sum_{j=1}^{m} H\left(y_j, F_2 \circ F_1\left(c_j, E_{text}\left(c_j\right)\right)\right) \quad , \tag{23}$$

where y_i is the one-hot vector where i-th instance is 1 otherwise is 0. $F_2 \circ F_1\left(c_j, E_{text}\left(c_j\right)\right)$ is represent the operation of phase3, but the model is the label context vector $c_j | j = 1...m$, and λ is the regularization penalty. The purpose of this restriction is to provide a reference point in space when a known label is given. After the label's embedded vector is input as a model, the network of the embedded layer and the classification layer should map the label to the label. The strongest point, avoiding the classification layer that has a tendency to have a large number of categories after L_s training.

S2. After the mapping relationship of the classification layer is established by using a large number of category data sets L_s, at least the quantity category data set L_u is migrated. In addition to sharing the classification weights, the structure of the embedded space is also migrated as domain knowledge, and the purpose is to make the model handle a small amount. In the case of a category sample, local adjustments can be made using a small number of categories of training using previously established structures, without the need to start training from the initial weights, thereby increasing the generalization of the model.

V. Experimental results

In this section, the environment configuration of the experiment in this paper will be described, and will present the data set of the experiment in the proposed method, furthermore, the measurement method verified on proposed method will discuss.

The experimental data set consist of 13,578 criminal law indictments and 43,422 criminal law judgments, the indictments and the judgments are most focused on crime of injury. Each indictment has the fact description of the case and the corresponding judgment has the judged law articles of the case. The cases of data set are involved 39 law articles, each law article has the law description. In order to compete with opponents, the experiment also verified the opponents data set. The opponents data set consist of 77,043 criminal law case, the format is the same as our data set. To verify the few shot classification performance, the data set is separated to high frequency set and low frequency set. The high frequency set contains the sample of number of occurrences is greater than 10, otherwise as low frequency.

The measurement verified on proposed model is used accuracy and f1 score, each metrics calculated by macro-average. The experiment is divided into two parts, first evaluated at the fully test set, and then evaluated at the low frequency set. The experiment compared three algorithms, including the proposed scheme with transfer, the scheme without transfer and the opponent's algorithm.

A. Multi-label Accuracy on Few-shot Set and Normal Set

Fig. 5 shows the experiment on multi-label data set (Collected by this paper) , the experiment verified the model on the normal set and rare category set, the opponents model was not support the multi-label classification, we modified the opponent's model for comparison. The multi-label set has a serious imbalance problem, limited by the difficulty of collection, so the proposed scheme performs slightly better than the opponent.

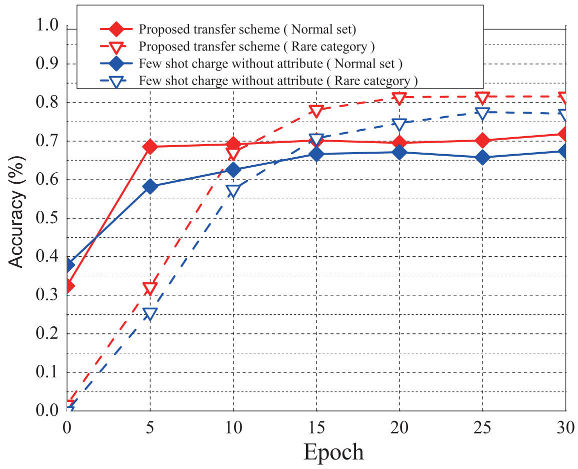


Fig. 5: Accuracy vs. per epoch.

B. Accuracy on Few-shot Set

Fig. 6 shows the experimental results of accuracy on the rare categories for the proposed schema with transfer, the proposed schema without transfer and the compared approach. The larger the epoch is, the more the data set training. The purpose of the experiment is to test the learning performance of different methods for a rare category of models. The experiment first trained the model on a high frequency analog set, then transfer to a rare amount of data and trained once. Then test the model's performance to classify a small number of categories. The rare number here is defined as the total number below 30. The Fig. 6 shows that the proposed scheme is better than the opponent in learning a small number of categories. Here, the opponent's model is without used the attribute tag, if used, the opponent will better then the proposed scheme.

Compared to the proposed scheme without weight shearing projection layer, the result shows that the projection layer can improves the performance of learning on few shot data. The regularization is slightly improves the performance on the scheme with weight shearing projection layer, for the scheme without projection layer the regularization has larger effect.

C. F1 Score on Few-shot Set

Fig. 7 shows the F1 score macro on the experiment. The F1 score can represent the classification performance of the model for each category. F1 indicates that the model has stronger recognition p for positive samples. Precision reflects the performance of the model to distinguish negative samples. The higher the precision, the stronger the distinguishing performance of the model for negative samples. F1-score is a combination of the two. As the Fig. 7, the experiment verified the model on our dataset and the opponent's dataset, both are in the case of few-shot data. As the result, the proposed scheme has the better recognition for few-shot categories in the opponent's dataset when the rare frequency defined as 30. The opponent's scheme will better then us when used the charge attribute. The result tested on our dataset shows that

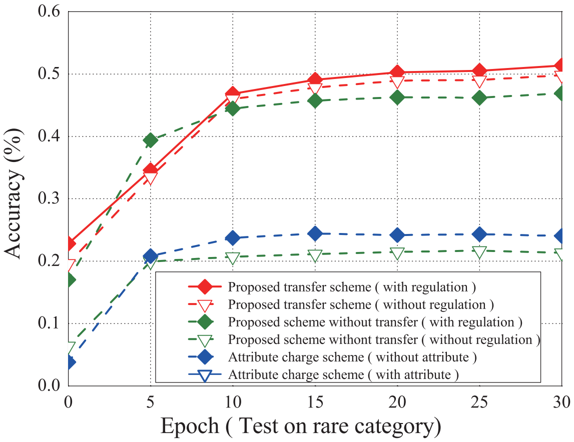


Fig. 6: Accuracy vs. per epoch.

the proposed scheme is better then the opponent when the opponent's scheme without used the attribute.

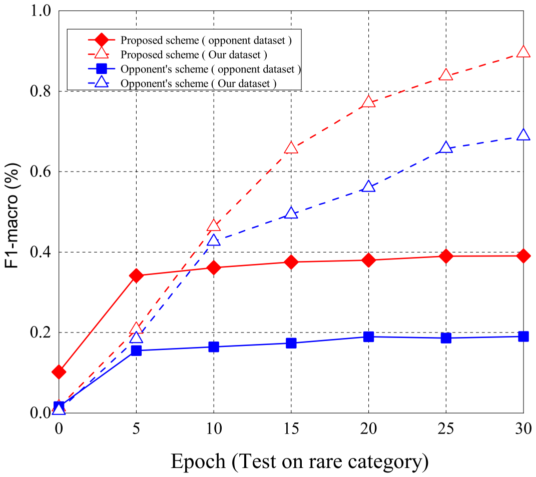


Fig. 7: F1 vs. per epoch.

VI. Conclusions

In this paper, we proposed a law article prediction scheme for legal case, There are four parts in our method: in the first part, a legal data set has collected; in the second part, a weighted approach based on text similarity is applied to produced the meaningful text representation of vector and the third part, a weight sharing projection layer is employed for prediction rare category sample, the last part used the prior knowledge of the projection layer transfer to the rare category set, the proposed scheme can achieves 49% of accuracy for prediction rare category which number is lower then 30.

VII. Acknowledgments

This work was supported by the National Science Council of Taiwan under Grant MOST 108-2221-E-305-005-.

References

[1] Yoon Kim. "Convolutional Neural Networks for Sentence Classification". In *Proceedings of Empirical Methods in Natural Language Processing* (EMNLPM 2017), pages 1746–1751, Doha, Qatar, Oct. 2014.

[2] G. Chen, D. Ye, Z. Xing, J.Chen, and C. Erik. "Ensemble Application of Convolutional and Recurrent Neural Networks for Multi-Label Text Categorization". In *Proceedings of International Joint Conference on Neural Networks* (IJCNN), pages 2377–2383, Anchorage, US, May. 2017.

[3] L. Torrey and J. Shavlik. "Transfer Learning". *Handbook of Research on Machine Learning Applications*, pages 23–34, January 2009.

[4] W. Wang, W. Vincent, and Z. Yu. "A Survey of Zero-Shot Learning: Settings, Methods, and Applications". *Transactions on Intelligent Systems and Technology*, 10(13):1–37, Feb 2019.

[5] B. Luo, Y.Feng, J. Xu, X. Zhang, and D. Zhao. "Learning to Predict Charges for Criminal Cases with Legal Basis". In *Empirical Methods in Natural Language Processing* (EMNLP), pages 2727–2736. Copenhagen, Denmark, Sep. 2017.

[6] C. H. Lampert, H. Nickisch, and S. Harmeling. "Attribute-Based Classification for Zero-Shot Visual Object Categorization". *IEEE Transactions on Pattern Analysis and Machine Intelligence* (TPAMI 2014), 36(3):453–465, July 2014.

[7] L. Shu, H. Xu, and B. Liu. " DOC: Deepopen Classification of Text Documents.". In *Proceedings of Empirical Methods in Natural Language Processing,* (EMNLP 2017), pages 2911–2916, Copenhagen, Denmark, Jan. 2017.

[8] B. E. Lauderdale and T. S. Clark. "The Supreme Court's Many Median Justices". *The American Political Science Review*, 106(4):847–866, Nov 2012.

[9] W. He, K. Liu, J. Liu, and Y. Lyu. "DuReader: A Chinese Machine Reading Comprehension Dataset from Real-World Applications". In *Proceedings of the Workshop on Machine Reading for Question Answering* (ACL 2017), page 3746, Melbourne, Australia, Jul. 2017.

[10] H. Zhong, Z. Guo, C. Tu, C. Xiao, and Z. Liu. "Legal Judgment Prediction via Topological Learning". In *Proceeding Association for Computational Linguistics* (ACL), page 35403549, Brussels, Belgium, Oct. 2018.

[11] S. Long, C. Tu, Z. Liu, and M. Sun. "Automatic Judgment Prediction via Legal Reading Comprehension". *Technical report from Beijing Tsinghua University*, September 2018.

[12] Z. Hu, X. Li, C. Tu, Z. Liu, and M. Sun. "Few-Shot Charge Prediction with Discriminative Legal Attributes". In *Proceeding of International Conference on Computational Linguistics* (COLING 2018), page 487498, New Mexico, USA, Aug. 2018.

[13] X. Jiang, H. Ye, Z. Luo, W. Chao, and W. Ma. "Interpretable Rationale Augmented Charge Prediction System". In *Proceedings of International Conference on Computational Linguistics: System Demonstrations* (COLING 2018), page 146151, Santa Fe, New Mexico, Aug. 2018.

[14] H. Ye, X. Jiang, Z. Luo, and W. Chao. "Interpretable Charge Predictions for Criminal Cases: Learning to Generate Court Views from Fact Descriptions". In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL:HLT 2018), pages 1854–1864, New Orleans, USA, Jun. 2018.

[15] B. Luo, Y. Feng, J. Xu, X. Zhang, and D. Zhao. "Learning to Predict Charges for Criminal Cases with Legal Basis". In *Proceedings of Conference on Empirical Methods in Natural Language Processing* (EMNLP 2017), page 27272736, Copenhagen, Denmark, Sep. 2017.

[16] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. "Distributed Representations of Words and Phrases and Their Compositionality". In *Proceeding of International Conference on Neural Information Processing Systems,* (NIPSL:HAT 2013), pages 3111–3119, Lake Tahoe, Dec. 2013.