Charge Prediction for Criminal Law with Semantic Attributes

Cong Zhou, Weipeng Cao*, and Zhiwu Xu*

College of Computer Science and Software Engineering, Shenzhen University, China
*Corresponding authors: {caoweipeng,xuzhiwu}@szu.edu.cn

Abstract. Most of the existing machine learning methods for charge prediction adopt the training mechanism of supervised learning. These algorithms have high requirements for the number of training samples corresponding to each crime. However, few or no cases are corresponding to some crimes in real scenarios, which leads to the poor performance of these models in practice. To alleviate this problem, we propose a novel Zero-Shot Learning (ZSL) based method for legal charge prediction tasks. Specifically, we define a set of semantic attributes to represent the domain knowledge of charges, which enables the model to migrate knowledge from seen charges to unseen charges. In this way, with the help of the ZSL mechanism, unseen charges and charges with a small number of training samples could be relatively predicted accurately. We evaluate the performance of the proposed method on a dataset collected from China Judgements Online, and the experimental results show that our method obtains 32.4% accuracy for the unseen charges and can largely retain the predictive power for the seen charges.

1 Introduction

In recent years, driven by emerging technologies such as big data, cloud computing, and the Internet-of-Things (IoT), application areas such as intelligent prediction of legal charges, Body Sensor Networks(BSNs) [1], tel-health systems [2], and real-time embedded systems [3] have all been rapidly developed. Among them, the intelligent prediction of legal charges can be considered as a classification problem based on case description and factual text data. A popular way to solve this type of problem is to use neural network algorithms, especially deep learning. For example, in 2017, Luo et al [4] provided an attention-based neural network to predict legal charges. Although neural network-based models have achieved promising accuracy on many benchmarks, there is a problem that cannot be ignored: most models need to collect a large number of cases corresponding to each crime during training. In other words, each criminal charge to be predicted requires enough training samples. However, in the real world, this premise is difficult to meet. For example, for some criminal charges such as “treason”, there are very few public cases. Even for many criminal charges, one cannot collect any public cases. This problem leads to the poor performance of

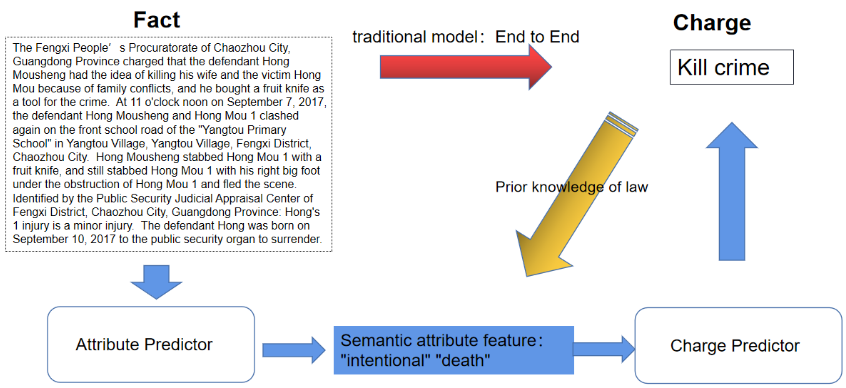

Fig. 1. An illustration of attribute-based charge prediction

many existing charge prediction models in practice. To solve the few-shot learning problem in the criminal charge prediction problem, Hu et al [5] applied the less-sample learning and provided a charge attribute based method to assist in charge prediction. However, this method still cannot effectively deal with the zero-shot learning problem.

To alleviate the above problem, we propose to employ Zero-Shot Learning (ZSL) technique [6] to improve the prediction accuracy of the model on those charges with limited training samples. The premise of using ZSL is that a reasonable connection can be established between the unseen classes and the seen classes so that the model can transfer the reasoning ability learned from the seen classes to the prediction of the unseen classes [7, 8]. For the problem studied in this paper, the first issue we have to solve is how to construct a semantic association among the majority criminal charges, minority criminal charges, and the criminal charges without any cases.

Therefore, according to the description of the provisions of the articles in the official documents and combined with the knowledge of the legal field, we define 19 semantic attributes (e.g., the manner of committing the crime, related objects, violence, and personal rights) for the criminal articles. These semantic attributes serve as a bridge between charges. With the help of the semantic attributes, we can use the ZSL technique to build an intelligent prediction model for legal charges. Specifically, using a large number of training samples for the seen classes, an intrinsic mapping model from the case fact text to attribute features can be obtained, which enables us to transfer the legal charge prediction from the text level to the attribute level, as shown in Fig 1. As there is still a data imbalance problem at the attribute level, we also introduce the synthetic minority over-sampling technique (SMOTE) [9] to improve the performance of the attribute predictor. Our main contributions are summarized as follows:

– We defined a set of semantic attributes for the charges of Criminal Law and constructed a ZSL-oriented dataset for criminal charge prediction.
– We proposed a ZSL/GZSL approach for legal charge prediction, focusing on the charges with fewer or even no training samples.
– We conducted several experiments on the self-constructed dataset, and the experimental results confirm the effectiveness of our semantic attributes and our charge prediction.

2 Related Work

2.1 Zero-Shot Learning

The concept of ZSL was first proposed as a problem in 2008 by Larochelle et al [10]. They formulated ZSL in general and used experiments to verify the feasibility and significance of ZSL in character recognition and multi-task sorting problems. In 2009, Lampert et al [11] first proposed to solve the problem of unseen class recognition by introducing attribute based classification - direct attribute prediction model (DAP) [12] and Indirect attribute prediction model (IAP) [13], and also proposed the "Animals with Attributes" dataset, which is nowadays the most common ZSL dataset in computer vision. In the same year, Farhadi et al [14] first transformed the class identification problem into an attribute description problem by generalizing attributes across classes and using the semantic information of attributes as a bridge between seen classes and unseen classes, thus enabling knowledge migration from seen classes to unseen classes. We recommend that readers refer to the review on ZSL for more progress in this field [6].

2.2 Charge Prediction

Although researchers in the legal field have always wanted to achieve intelligent legal judgments by machines, the methods proposed in the early years were quite limited due to technical limitations. It was only after the rapid development of machine learning and its successful application in several fields that legal charge prediction was gradually improved and made possible. In 2012, Lin et al [15] extracted 21 feature labels for this purpose and designed a machine learning model for classification. In recent years, researchers have suggested combining neural networks with natural language processing. In 2017, Luo et al [16] provided an attention-based neural network approach to train a charge prediction model. However, the above modes are built on charges with sufficient training data, which is ineffective in the charges with fewer samples. In 2018, to improve the performance on the the charges with fewer samples, Hu et al [5] provide a method based on the attributes of the charges for charge prediction. The method selected 10 representative attribute characteristics, including violence, profitable purpose, buying, and selling. But this method is still limited to focusing attention on supervised learning and almost ineffective in unseen charges.

3 Approach

In this section, we present our approach for charge prediction. We first define a set of semantic attributes to represent charge classes, and then propose a charge prediction method based on the semantic attributes.

3.1 Semantic Attributes of Offence

In ZSL, the classes that appear in the training phase are called seen classes while the ones that do not appear in the training phase are called unseen classes. To predict the unseen classes, one needs to express information of the unseen classes in terms of the one of seen class. Semantic Attributes are one of the effective way to bridge between seen and unseen classes as well as achieve knowledge migration. In this section, we define a set of semantic attributes , according to the description of the offences of criminal law.

To start with, we introduce the principles for defining of the semantic attributes:

1. Semantic attributes should be derived from the descriptions of each charge.
2. Semantic attributes should be able to distinguish different charges.
3. Most of semantic attributes should be common (*i.e.*, involved to a good number of charges, including the seen and unseen charges).

Clearly, these principles enable the knowledge migration between charges.

Guided by these principles, we carefully study the Criminal Law of the People's Republic of China, which consists of ten chapters and four hundred and fifty-one legal provisions in total. Then we find out the legal provisions related to the crime, and represent them as crime-provision pairs (crime, provision description). For example, one of the pair is (kill crime, intentional homicide), representing that the conditions for forming the Kill crime are *intentional* and causing *death* of the victim. Another example is (negligent homicide crime, negligence causing death), whose crime is quite closed to kill crime but its description emphasizes the condition of *negligence*, instead of *intentional*. This demonstrates our semantic attributes could help us to distinguish different (but similar) crimes. Some of these attributes are common to most crimes, such as *death* appear in both kill crime and negligent homicide crime. Moreover, some common attributes can be used as a necessary condition for the crime, such as the kill crime must satisfy the *death* condition. In addition, some (values of the) attributes are very representative so that then can be used as a sufficient condition. For example, if the location for crime is an ancient tomb, then clearly the crime is the one of excavation of ancient cultural sites. Such attributes can clearly distinguish the corresponding crimes from other ones.

Finally, according to logical perceptions, we grouped the extracted attributes into several classes, which contributes to making the attribute more general. For example, *intentional* and *negligence* can be grouped into *intent*. And both *death* and *detention* involve personal rights, so we group them into *personal rights*.

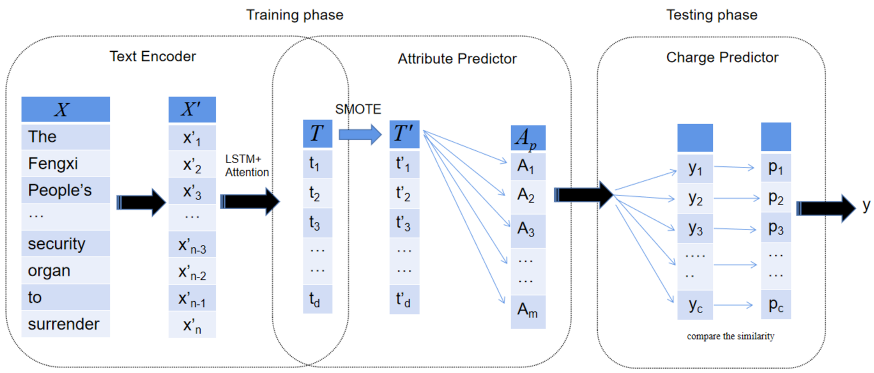

Fig. 2. Framework of Charge Prediction

Other attribute categories include, for example, consequences, location, violence, etc. In total, we define 19 semantic attributes and 153 features (*i.e.*, attribute values) in total.

3.2 Charge Prediction

In this section, we present our approach for charge prediction, based on semantic attributes we defined for the criminal law. The key idea of our approach is to reduce the task of charge prediction into the one of attribute prediction. Figure 2 shows the framework of our approach, which consists of three components, namely, *text encoder*, *attribute predictor*, and *charge predictor*. *Text encoder* encodes the factual text of the case into a multi-dimensional vector word by word, according to a pre-trained corpus; fed by the multidimensional vector from *text encoder*, *attribute predictor* predicts the semantic attributes for the case, and finally, *charge predictor* output a charge according to the predicted semantic attributes.

To start with, we give a formal definition of charge prediction. The charge prediction task is defined as a function (or model) $M : X \to Y$, where X denotes the set of the given cases and Y denotes the set of charges. Let Y_s denote the set of seen charges and Y_u denote the set of unseen charges. Thus, we have $Y = Y_u$ and $Y = Y_u \cup Y_s$ for ZSL and GZSL, respectively. Moreover, a sample data in charge prediction is represented as (x, y), where $x \in X$ is the factual text of a case and represented as its word sequence $x = [x_1, x_2, \dots, x_n]$, n is the length of the factual text, and $y \in Y$ is the true charge of the case. Let W denote the set of words appearing in X. We assume there is a pre-trained corpus for W.

Text Encoder According to the pre-trained corpus of W, each word x_i of a case x is first converted into a multi-Dimensional word vector x'_i, where $x'_i \in R^{wd}$ and wd is dimension of the word vector. Then the resulting word vector sequence

$x' = [x'_1, x'_2, \ldots, x'_n]$ is fed into a neural network to extract semantic information from the text, wherein Long Short-Term Memory (LSTM) and Self-Attention Mechanism are used. The result is a vector $T = [t_1, t_2, \ldots, t_d]$ representing the semantic information of the text, where d is the text vector dimension. Let $E : X \to R^d$ denote the text encoder.

Attribute Predictor To achieve knowledge migration between seen charges and unseen charges, we define a set of semantic attributes for charges in Section 3.1. Based on this set, we represent each charge y_i as a vector of semantic attributes, that is, $y_i = [a_1, a_2, \ldots, a_m]$, where $y_i \in Y$, a_i is a semantic attribute, and m is the number of semantic attributes. According to Principle 2, the vector of each charge is different. As mentioned before, we reduce the the original text-to-crime prediction task into the text-to-attribute prediction one. Formally, the attribute prediction task is defined as a function $F : T \to [A_1, A_2, \ldots, A_m]$, where T is the set of possible text vectors and A_i is the set of features for the i-th semantic attribute.

Given a text vector T, the attribute predictor for each attribute is trained by sampling the true charges $\hat{y}$. However, although we consider the attributes that are as common as possible, there are still some attributes that only appear in several similar charges, due to that they are critical to these similar charges. This could yield the problem that the distribution of samples for different attributes is extremely unbalanced, resulting in a model tending to favor the charge with more attribute samples. Therefore, we have to consider the attributes with fewer samples. For that, we introduce the SMOTE algorithm to expand the original training set T to a set T' containing the original and generated samples, such that the distribution of samples for each attributes is relatively balanced. The basic idea of the SMOTE algorithm is to analyze the samples of the minority classes and artificially synthesize new samples and add them to the dataset.

Once the attribute predictors are trained, we are able to predict the semantic attributes for the factual text for a given case. In detail, given a case x, the predict attribute vector for x is $A_x = F(E(x)) = [a_1, a_2, \ldots, a_m]$, where $a_i \in A_i$ is predicted by the i-attribute predictor.

Charge Predictor Our goal is to predict a charge. Thus, after predicting an attribute vector $A_x = [a_1, a_2, \ldots, a_m]$ for a given case x, we compare the attribute vector A_x with the predefined attribute vector A_i of each charge y_i, and take the most similar one as the predicted result. Here we use the cosine similarity, which is computed as follows:

$$p_x = \frac{A_x \cdot A_i}{\sqrt{|A_x| * |A_i|}} \tag{1}$$

where p_x denotes the similarity between the predicted attribute vector A_x and the i-attribute vector A_i. And the final predicted charge

$$y_x = y_{h(max(p_1, p_2, \ldots, p_c))} \tag{2}$$

where $h(max(p_1, p_2, \ldots, p_c))$ is a function that returns the index for the maximum value and c denotes the number of charges.

Finally, as a notice, the time complexities of our approach in the training phase and the prediction phase are respectively $O(m \times |X|)$ and $O(m + |Y|)$, where m is the number of attributes.

4 Experiments

We construct a ZSL-oriented dataset from *China Judgements Online*[1] (Section 4.1), following the way in Section 3.1. With this dataset, we evaluate the performance of our proposed ZSL charge prediction model (Section 4.2). We also perform ablation analysis experiments on the attention mechanism and the SMOTE technique (Section 4.3).

4.1 Dataset and Criteria

In our experiments, we collected 559,830 samples of the Criminal Law of the People's Republic of China from *China Judgements Online*, covering 91 charge classes. The charge classes with less than 100 samples are considered as unseen classes, which have 31 in total. All the samples of unseen classes are used as the testing samples for both ZSL and GZSL. On the contrary, the remaining charges are treated as seen classes. Most of their samples are used as the training samples for ZSL and GZSL, and the remaining samples are used as the testing samples for GZSL, so the testing set of GZSL contains the remaining samples from seen classes and the samples from the unseen classes. As mentioned in Section 3.1, the attribute vector is composed of 19 attributes and 153 features. The statistics of the dataset are given in Table 1.

Table 1. Statistics of the dataset.

Dataset	**ZSL**	**GZSL**
Charge	31 (unseen classes)/91	
Attribute	19	
Train samples	485142	
Test samples	1328	74688

In our experiment, we use the accuracy Acc to evaluate the prediction results. In particular, we use Acc_s and Acc_u to denote the accuracy of seen classes and the accuracy of unseen classes, respectively. As our dataset is an extremely unevenly distributed one, we also use the harmonic average Acc_h of Acc_s and Acc_u

$$Acc_h = \frac{2 * Acc_s * Acc_u}{Acc_s + Acc_u} \tag{3}$$

[1] `https://wenshu.court.gov.cn/`

4.2 Charge Prediction Experiment

The charge prediction results of our model are given in Table 2. For comparison, Table 2 also includes the results of the traditional supervised learning, that is, SVM, KNN, CNN, and LSTM. Although there are no relevant unseen class samples during the training process, the accuracy of our model on the test set containing the samples of 31 unseen classes (*i.e.*, the test set for ZSL) is 32.4%, which is much larger than the random distribution probability 3.2% (1/31). This indicates that our proposed semantic attribute set is effective in knowledge migration from seen to unseen classes.

Table 2. The accuracies of charge prediction for various models.

Model	Acc_u	Acc_s	Acc	Acc_h
SVM	-	82.8	81.3 (-1.5)	-
KNN	-	86.9	85.4 (-1.5)	-
CNN	-	88.2	86.6 (-1.6)	-
LSTM	-	**89.3**	**87.7 (-1.6)**	-
Our model	**32.4**	79.8	**79.0 (-0.8)**	**46.1**

On the testing set for GZSL, our model achieves an accuracy of 79.0%, wherein the accuracy for the samples from seen classes is 79.8%. Compared to traditional supervised learning, although our result is a little degraded, our model has the smallest reduction in overall accuracy. In other words, our model not only achieves performance on charge prediction closed to the traditional supervised learning but also guarantees a good accuracy (over 30% in our experiment) for the unseen charges. Moreover, the harmonic average of Acc_u and Acc_s for our model is 46.1%. Note that our GZSL test set contains 74,688 test samples, wherein less than 1.7% (1,328/74,688) samples are from 34.1% (31/91) unseen classes. This indicates that, unlike traditional supervised learning, our model would not focus on the classes with a large number of samples to get higher overall performance. The above results demonstrate that our semantic attributes are suitable for expressing charges, that is, they are not only logically helpful for better understanding but also able to achieve a closed expressiveness to the one of the abstract features extracted by machine learning itself.

4.3 Ablation Experiments

To evaluate the effectiveness of the attention mechanism and the SMOTE algorithm, we design separate ablation experiments. In detail, we remove the attention mechanism and the SMOTE algorithm from the text encoder and the attribute predictor, respectively. The results of ablation experiments are presented in Table 3.

As shown in Table 3, we can observe that all the accuracies decrease if either the attention mechanism or the SMOTE algorithm is removed. Moreover, we

Table 3. Results of ablation experiments

Model	Acc_u	Acc_s	Acc	Acc_h
Our model	**32.4**	**79.8**	**79.0**	**46.1**
W/o attention	30.5	75.1	74.3	43.4
W/o smote	25.8	77.3	76.4	38.7

also found that the performance impact of the attention mechanism on charge prediction is smaller than the one of the SMOTE algorithm. This is because the attention mechanism is only helpful to filter the semantic information in the text embedding process, while the SMOTE algorithm is effective to solve the model over-fitting problem caused by data imbalance, which is the key to our model.

5 Conclusion

In this work, we have defined a set of semantic attributes for Criminal Law, which can achieve knowledge migration from seen charges to unseen charges. We have also proposed a ZSL/GZSL approach for charge prediction, based on the semantic attributes, which provides a feasible solution for the data-driven model to predict crimes that are difficult to collect precedents. Experimental results on our ZSL-oriented dataset have shown that our approach not only achieves comparable accuracy to the traditional supervised learning on seen charges but also achieves a prediction accuracy of over 30% for the unseen charges.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (61836005 and 62106150), Guangdong Basic and Applied Basic Research Foundation (2019A1515011577), Stable Support Programs of Shenzhen City (20200810150421002), CCF-NSFOCUS (2021001), and CAAC Key Laboratory of Civil Aviation Wide Survellence and Safety Operation Management & Control Technology (202102).

References

1. Longfei Qiu, Keke Gai, and Meikang Qiu. Optimal big data sharing approach for tele-health in cloud computing. In *2016 IEEE International Conference on Smart Cloud, SmartCloud 2016, New York, NY, USA, November 18-20, 2016*, pages 184–189. IEEE Computer Society, 2016.
2. Han Qiu, Meikang Qiu, and Zhihui Lu. Selective encryption on ecg data in body sensor network based on supervised machine learning. *Information Fusion*, 55:59–67, 2020.

3. Meikang Qiu, Chun Xue, Zili Shao, and Edwin Hsing-Mean Sha. Energy minimization with soft real-time and DVS for uniprocessor and multiprocessor embedded systems. In Rudy Lauwereins and Jan Madsen, editors, *2007 Design, Automation and Test in Europe Conference and Exposition, DATE 2007, Nice, France, April 16-20, 2007*, pages 1641–1646. EDA Consortium, San Jose, CA, USA, 2007.
4. Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. Learning to predict charges for criminal cases with legal basis. *CoRR*, abs/1707.09168, 2017.
5. Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. Few-shot charge prediction with discriminative legal attributes. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 487–498, 2018.
6. Weipeng Cao, Cong Zhou, Yuhao Wu, Zhong Ming, Zhiwu Xu, and Jiyong Zhang. Research progress of zero-shot learning beyond computer vision. In *International Conference on Algorithms and Architectures for Parallel Processing*, pages 538–551. Springer, 2020.
7. Zhongwu Xie, Weipeng Cao, and Zhong Ming. A further study on biologically inspired feature enhancement in zero-shot learning. *International Journal of Machine Learning and Cybernetics*, 12(1):257–269, 2021.
8. Yuxuan Luo, Xizhao Wang, and Weipeng Cao. A novel dataset-specific feature extractor for zero-shot learning. *Neurocomputing*, 391:74–82, 2020.
9. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, 16:321–357, 2002.
10. Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (AAAI 2008)*, pages 646–651, 2008.
11. Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pages 951–958, 2009.
12. Kongming Liang, Hong Chang, Shiguang Shan, and Xilin Chen. A unified multiplicative framework for attribute learning. In *IEEE International Conference on Computer Vision (ICCV 2015)*, pages 2506–2514, 2015.
13. Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(3):453–465, 2014.
14. Ali Farhadi, Ian Endres, Derek Hoiem, and David A. Forsyth. Describing objects by their attributes. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pages 1778–1785. IEEE Computer Society, 2009.
15. Wan-Chen Lin, Tsung-Ting Kuo, Tung-Jia Chang, Chueh-An Yen, Chao-Ju Chen, and Shou-de Lin. Exploiting machine learning models for chinese legal documents labeling, case classification, and sentencing prediction (in chinese). *Int. J. Comput. Linguistics Chin. Lang. Process.*, 17(4), 2012.
16. Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. Learning to predict charges for criminal cases with legal basis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 2727–2736, 2017.