Label Definitions Augmented Interaction Model for Legal Charge Prediction

Liangyi Kang[1,2], Jie Liu[1,2(✉)], Lingqiao Liu[3(✉)], and Dan Ye[1,2]

[1] State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing, China
{kangliangyi15,ljie,yedan}@otcaix.iscas.ac.cn
[2] University of Chinese Academy of Sciences, Beijing, China
[3] University of Adelaide, Adelaide, Australia
lingqiao.liu@adelaide.edu.au

Abstract. Charge prediction, determining charges for cases by analyzing the textual fact descriptions, is a fundamental technology in legal information retrieval systems. In practice, the fact descriptions could exhibit a significant intra-class variation due to factors like non-normative use of language by different users, which makes the prediction task very challenging, especially for charge classes with too few samples to cover the expression variation. In this work, we explore to use the charge (label) definitions to alleviate this issue. The key idea is that the expressions in a fact description should have corresponding formal terms in label definitions, and those terms are shared across classes and could account for the diversity in the fact descriptions. Thus, we propose to create auxiliary fact representations from charge definitions to augment fact descriptions representation. Specifically, we design label definitions augmented interaction model, where fact description interacts with the relevant charge definitions and terms in those definitions by a sentence- and word-level attention scheme, to generated auxiliary representations. Experimental results on two datasets show that our model achieves significant improvement than baselines, especially for dataset with few samples.

Keywords: Legal charge prediction · Label definitions · Interaction model · Auxiliary representation · Augmented fact representation

1 Introduction

The task of charge prediction is to determine appropriate charges, such as *theft* or *robbery*, for given cases by analyzing the textual fact descriptions. Automating charge prediction technology could be practically useful for online legal assistant systems, which provide legal consulting for users in a cost-effective way.

In practice, users have different writing habits while inputting the fact of cases. Fact descriptions comprise a substantial amount of diverse non-normative use of language. For example, the cases of robbery in Fig. 1 all involve "theft",

D. Hiemstra et al. (Eds.): ECIR 2021, LNCS 12656, pp. 270–283, 2021.
https://doi.org/10.1007/978-3-030-72113-8_18

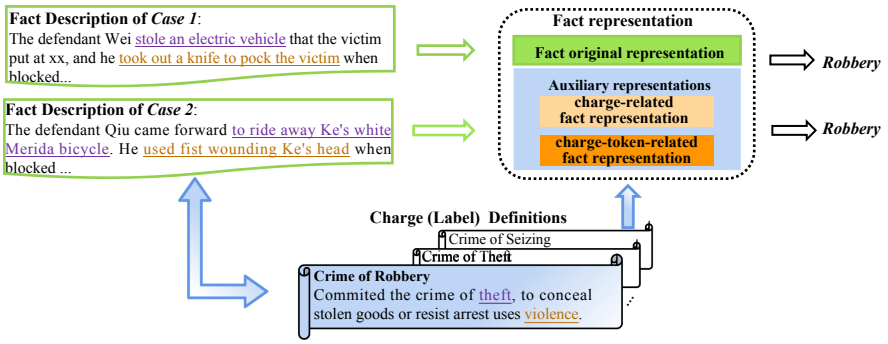

Fig. 1. Illustration of our method. Green boxes are two robbery case descriptions and the blue box contains label definitions–charge definitions. The related charges are identified (indicated by the blue double arrow) via sentence-level attention and aggregated to create the auxiliary representation I, charge-related fact representation. Then key words in cases align to terms in identified charge definitions via word-level attention (aligned words are labeled by the same color), which are then formed as the auxiliary representation II, charge-token-related fact representation. The two auxiliary representations combine with original fact representation to predict the label–robbery. (Color figure online)

but the legal term "theft" may be implicitly expressed like *"stole an electric vehicle"* or *"came forward to ride away Ke's white Merida bicycle"*. Consequently, the representation of fact descriptions may exhibit considerable intra-class variation which may lead to prediction failure at the test stage. This could be more pronounced for charge classes with only a few examples since the samples are not sufficient for learning a predictive model robust to expression variation.

To address this issue, we introduce label definitions, the charge definition, to create more robust fact representations for charge prediction. We propose to create auxiliary fact representations from the charge definitions to augment the fact representation. Those auxiliary representations are essentially projections of the fact description in the semantic space of charge definitions. Our motivation is that the expressions in a fact description should have corresponding formal terms in label definitions, and those formal terms can provide an alternative view of the expressions in fact description. Note that many of those formal terms are shared across charge classes and are less diverse. Thus, using elements in charge definitions to re-interpret fact description and generate auxiliary representations could have the potential to account for the diversity in the fact description.

Specifically, we design a label definitions augmented interaction model integrating sentence- and word-level attention to generate two auxiliary fact representations. We identify the related charge definitions through sentence-level attention between fact description and charge definitions, and then aggregate the holistic features of relevant charge definitions to create the first auxiliary representation, named as charge-related fact representation. The relevant charge definitions identified in the course of producing the first auxiliary representation will also serve for creating the second auxiliary representation. To create

the second representation, we further consider finer-grained word-level attention between the fact description and related charge definitions. Relevant words from relevant charge definitions are attended and aggregated through a recurrent neural network to generate the second auxiliary representation, named as charge-token-related fact representation. We illustrate our model by an example in Fig. 1. Case 1 and case 2 in Fig. 1 belong to the same class, *robbery*, but with different expressions. With the proposed method, they will be firstly related to the charge definition of *robbery*. Then the statements of *"stole an electric vehicle"* and *"took out a knife to poke the victim"* in case 1, *"came forward to ride away Ke's white Merida bicycle"* and *"used fist wounding Ke's head"* in case 2 will be softly aligned to the terms "theft" and "use violence" in *robbery* definition through interaction. By reinterpreting the fact descriptions through aligned terms, those two cases become more similar. The final charge prediction is based on the original and auxiliary fact representations, and one can expect the prediction made on this fact representation will be more robust.

To investigate the advantage of our method on charge prediction, we conduct experiments on real-world datasets. Experimental results show that our model outperforms baselines, especially on dataset with few samples. We also conduct ablation studies to analyze the effectiveness of each component in our model, and visualize the impact of introducing charge definitions.

2 Related Works

Charge prediction focuses on learning representation of fact descriptions and feeding them into classifiers to make the judgment. At the early stage, [13–15,18] attempt to extract shallow text features from fact descriptions or create handcrafted features to represent fact descriptions, which are hard to generalize to large datasets due to the diverse expression of fact descriptions. Inspired by the success of deep learning, [8,16,26,27] employ neural models with external information to capture the high-level semantic information. [16] use a separate two-stage scheme to extract the related articles and then attend them attentively to fact representation. [8] design 10 legal attributes to help the few-shot charges prediction. They both need a large amount of feature engineering, either design features or relations between subtasks. LJP [27] and MPBFN [25] model multiple legal subtasks by multi-task learning framework to assist prediction. LegalAtt [2] uses law article to perfect fact representation. However, one article may include more than one charges, which could obscures the fact representation. Instead, we augment fact representation to assist charge prediction by creating auxiliary representation from charge definitions by an interaction model.

Our model is also related to attention and memory in deep learning [1,6,17,19,20,22]. Although researchers propose various neural architectures with memory and attention for NLP problems [7,12,21], they either only consider sentence-level or only word-level alignment between sentences. In contrast, we combine them jointly to form auxiliary representation, where sentence-level interaction identifies relevant charges, and a finer-grained word-level interaction on the top of identified charge definitions makes the generated fact representation more robust.

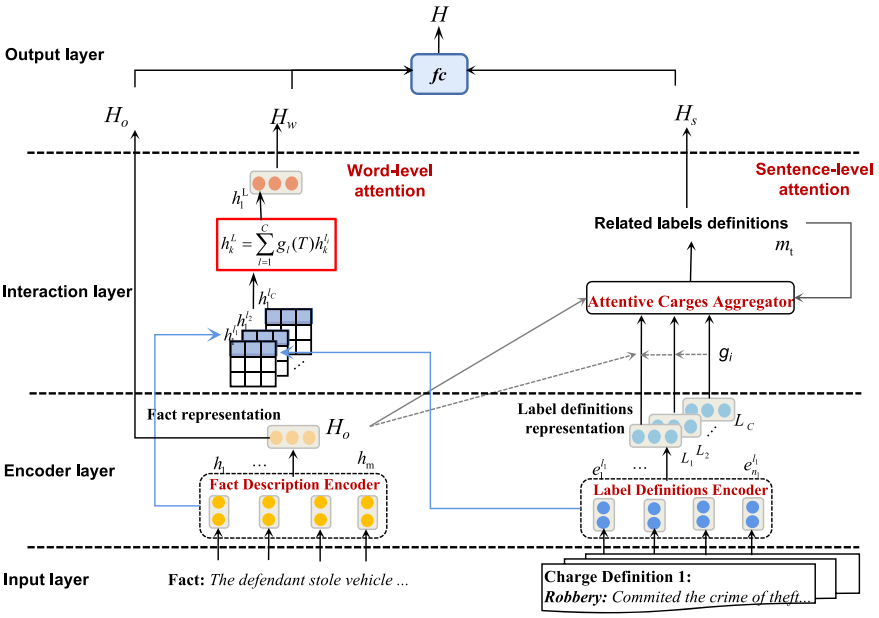

Fig. 2. The architecture of our model. Fact description encoder embeds the fact description into the original fact representation H_o. Sentence-level attention creates auxiliary representation I: attentive charges aggregator is iteratively to identify related charges that are then aggregated to generate H_s. On top of identified charges, word-level attention creates auxiliary representation II: each word in a fact description is represented by the combination of the terms in related charge definitions. The combined intermediate representations are aggregated through a GRU to generate H_w. At last, H_o, H_s and H_w are concatenated to form final fact representation H for prediction.

3 The Proposed Model

3.1 Problem Formulation

Charge prediction is to predict the corresponding charges l for a given fact description d, where fact description d consists of a sequence of words $\{w_1^d, w_2^d, \cdots, w_m^d\}$, and its label is a C dimensional multi-hot vector – a fact description may correspond to one or multiple labels in C classes. The charge definition for the i-th label l_i is a sequence of words $\{w_1^{l_i}, w_2^{l_i}, \cdots, w_{n_i}^{l_i}\}$.

3.2 Framework

To generate a robust fact representation for prediction, we propose a label definitions augmented interaction model integrating sentence- and word-level attention. The architecture is shown in Fig. 2. The final fact representation H is the concatenation of three representations: 1) the original fact representation (H_o), 2) the auxiliary representation I, charge-related fact representation (H_s), 3) the auxiliary representation II, charge-token-related fact representation (H_w).

3.3 Fact Description Encoder

Giving a fact description with a sequence of word embeddings, we use Gated Recurrent Unite [3] to encode contextual information of each word.

$$h_i = GRU(w_i^d, h_{i-1}), \tag{1}$$

where h_i is the hidden state of the GRU at time step i.

For a fact description, the words and consequently those hidden variables do not contribute equally to convey the semantic meaning of a text, and long fact description will involve many less informative words. To suppress the negative impact of the non-informative words, we use attention mechanism to assign each hidden state an importance weight α_i.

$$\alpha_i = softmax(W_2 tanh(W_1 {h_i}^T)), \tag{2}$$

where $\alpha_i \in [0, 1]$ is the weight of h_i and $\sum_i \alpha_i = 1$. W_1 and W_2 are trainable parameters. The holistic representation of original fact description H_o is computed as a weighted sum of those hidden variables:

$$H_o = \sum_{i=1}^{m} \alpha_i h_i. \tag{3}$$

3.4 Charge Definitions Encoder

Each class label l_i is associated with a charge definition. For each charge definition, we use the same CNN [9] to encode the sequence of n words into a sequence of vectors. Since we will deal with a large number of labels, using CNN gives us better training efficiency than using GRUs.

$$e_j^{l_i} = \mathbf{CNN}(w^{l_i}_{j-\frac{s-1}{2}}, \cdots, w^{l_i}_{j+\frac{s-1}{2}}), \tag{4}$$

where the window size of CNN is s. Then we sum up these vectors to create the holistic representation of each charge definition.

$$L_i = \sum_{j=1}^{n_i} e_j^{l_i}. \tag{5}$$

3.5 Two Auxiliary Fact Representations from Charge Definitions

The first auxiliary fact representation is created through the sentence-level attention between the fact description and charge definitions. Its creation process iterates between two steps: identifying related charges and attentively aggregating the holistic representation of related charge definitions. After those iterations, relatedness weights of each charge will be obtained and they will also be used as the basis for creating the second auxiliary fact representation. The second auxiliary fact representation is generated from word-level attention, which aligns terms in charge definitions with the expressions in the fact description and aggregates those terms through a recurrent neural network. We elaborate the creation of those two auxiliary representations as follows.

Auxiliary Representation I: Charge-Related Fact Representation Created via Sentence-Level Attention Related Charges Identification. Identifying related charges is realized by calculating an attention weight for each charge to indicate the relatedness. Specifically, we exploit episodic memory attention mechanism [24] to iteratively calculate the attention weight from the correlation between the charge definitions and fact description and memory m_t, where m_t can be seen as the summary of already identified charges up to the t-th iteration and will be updated at each iteration. With more iterations, the unrelated charges can be filtered out. The memory m_t is initialized with original holistic representation of fact description, that is, $m_0 = H_o$.

Formally, we use following formulas to calculate the attention weight g of each charge definition at the t-th iteration.

$$z_i = [L_i \circ H_o; L_i \circ m_t; |L_i - H_o|; |L_i - m_t|], \tag{6}$$

$$g_i(t) = softmax(W_2^a tanh(W_1^a z_i)), \tag{7}$$

where $\circ$ is the element-wise product, $|.|$ is the element-wise absolute value, and; represents concatenation of the vectors. W_1^a and W_2^a are trainable parameters.

Attentive Charge Aggregator. Once the attention weight of each charge is calculated, we update the memory by performing weighted summation over charge definition representations.

$$m_{t+1} = \sum_{i=1}^{C} g_i(t) L_i. \tag{8}$$

Finally, we concatenate original fact representation with the last memory and the previous memory, and feed them into a fully-connected layer to create charge-related fact representation by using the following equation:

$$H_s = fc([H_o; m_T; m_{T-1}]), \tag{9}$$

where fc denotes the fully connected layer.

Auxiliary Representation II: Charge-Token-Related Fact Representation Created via Word-Level Attention

In the course of creating the above representation, both fact description and charge definitions are represented by holistic feature vectors. In other words, the interaction between fact and charge definitions is only at the sentence level. The second auxiliary representation steps further introducing interaction at the word level. Specifically, for each hidden variable h_k in the fact description, we first compute its matching score towards each word $e_j^{l_i}$ in each charge definition l_i by inner-product. Then $e_j^{l_i}$ is attentively aggregated to an intermediate representation $h_k^{l_i}$:

$$\beta_j = softmax(h_k \cdot {e_j^{l_i}}^T), \tag{10}$$

$$h_k^{l_i} = \sum_{j=1}^{n_i} \beta_j e_j^{l_i}. \tag{11}$$

The above intermediate representation is defined w.r.t. each charge definition l_i. In our method, we further perform a weighted summation over $h_k^{l_i}$ for different charge definition l_i. The weight is the attention weight $g_i(T)$ calculated at the last iteration T in Eq. (7). Using this weight fits our intuition that the terms in the related charges are more relevant to the expressions in the fact description.

$$h_k^L = \sum_{i=1}^{C} g_i(T) h_k^{l_i}. \tag{12}$$

Note that h_k^L can be viewed as a projection of h_k in the space spanned by $e_j^{l_i}$.

After obtaining h_k^L for each word in the fact description, we process the sequence by a new GRU and obtain the last hidden state $\bar{h}_l$:

$$\bar{h}_t = GRU(h_t^L, \bar{h}_{t-1}). \tag{13}$$

We concatenate original and the projected fact representation, and feed them into a fully-connected layer to generate charge-token-related fact representation.

$$H_w = fc([H_o; \bar{h}_l]). \tag{14}$$

3.6 The Output

Finally, we concatenate all the generated representations and feed them into a fully-connected layer to generate the final fact representation H.

$$H = fc([H_o; H_s; H_w]). \tag{15}$$

Since the evaluated tasks are multi-label problem, we input H into a linear classifier layer with sigmoid activation function to predict the probability, p_{il}, of each labels. The loss function for training is as follows:

$$Loss = -\frac{1}{N} \sum_{i=1}^{N} \sum_{l=1}^{C} [y_{il} log(p_{il}) + (1 - y_{il}) log(1 - p_{il})], \tag{16}$$

where N is the number of training data, C is the number of labels. $y_{il} \in \{0, 1\}$ is the original output of l-th class for i-th training sample and p_{il} is the estimated likelihood of the l-th label being true.

Table 1. Statistics of datasets.

Datasets	Training samples	Validation samples	Test samples	Charge classes
CAIL150k	154592	17131	32500	202
CAIL30k	32506	17131	32500	168

4 Experiments

4.1 Datasets

Table 1 shows the statistics of our used datasets. We use publicly available datasets of the Chinese AI and Law challenge (CAIL2018)[1][23]: `CAIL150k` dataset and `CAIL30k` dataset. `CAIL150k` and `CAIL30k` are different scales with 150,000 and 30,000 training samples respectively[2]. It is worth noting that in these two datasets the distribution of charges is quite imbalanced. In `CAIL150k`, the 31% charges in the training set have less than 100 cases, taking up only 1.88% of the total number of cases. In `CAIL30k`, 42% charges have less than 10 cases, taking up only 0.89% of the total number of cases.

As for charge definitions, they are extracted from articles in the Criminal Law of the People's Republic of China. Specifically, in criminal law, except for articles irrelevant to specific charges, each article may include more than one charges, their corresponding charge definitions, and punishment. We merge the charge definitions scattered in multiple articles. A snippet of cases and charge definitions is illustrated in Fig. 1.

Evaluation Metrics. We employ accuracy (Acc.), macro-precision (MP), macro-recall (MR) and macro-F1 (MF1) as evaluation metrics. Macro-precision/recall/F1 are calculated by averaging the precision, recall, and F1 of each class, which are metrics commonly used for multi-label classification tasks. The experimental results on test set use the parameters providing the best validation performance.

4.2 Training Setup

As all the sentences in charge definitions and fact descriptions are written in Chinese without word segmenting, we apply jieba[3] for word cut. We set the maximum length of fact description to 500, charge definitions to 110. We use pre-trained GloVe [5] vectors as our initial word embeddings. In practice, we choose the 64 dimensional embedding vectors trained on `baidubaike`. The iteration time

[1] http://cail.cipsc.org.cn/index.html.

[2] In CAIL2018 dataset, CAIL150k is ./exercise_contest/data_train.json. CAIL30k is ./final_test.json. They share the same validation and test set (./exercise_contest/data_valid.json and data_test.json).

[3] https://github.com/fxsjy/jieba.

T in Eq. (12) is set as 3. Adam [10] is used as the optimizer and the learning rate is initialized as 0.005 and halved in every other epoch. The epoch size is 20.

4.3 Baselines

We compare our model against several text classification models and existing charge prediction methods, where we only consider the methods with no feature engineering. They can be categorized into four categories:

- **Not using charge definitions for classification.** We implement deep learning models, such as multi-layers Convolution Neural Network (**CNN_classify**) [9], Gated Recurrent Unite (**GRU_classify**) [3] and **BERT** [4] for fact representation learning and classification.
- **Matching the fact representation with charge definitions for classification.** We train a **Siamese CNN** [11] to match the representations of fact description and charge definitions to find the best matched labels.
- **Augmenting fact description with charge definitions for classification.** We implement **Fact-Law AN** [16] that uses relevant law articles, selected by SVMs, to serve as a legal basis for encoding the fact description. To demonstrate the advantage of our model in considering sentence- and word-level interaction jointly, we implement improved memory network (**MemNet**) [12] and **GA_Reader** [21], which employ multi-iterative interaction between query and document at sentence- and word-level respectively for question-answer task.
- **Using multi-task learning for classification.** We re-implement existing charge prediction models **TopJudge** [27] and **LegalAtt** [2], which introduce related legal tasks to train a better fact representation in multi-task mode.

4.4 Results

Experimental results on two scale datasets are shown in Table 2. The observations are as followings:

- Generally speaking, models without incorporating charge definitions (**CNN_classify**, **GRU_classify**) perform inferior to their charge-definition-incorporated counterparts. **BERT** works better due to its strong pre-trained model. This observation clearly demonstrates the benefit of introducing label definitions.
- Incorporating charge definitions through matching based approaches (**Siamese CNN**) works to some extent, although their performance is still worse than methods using more sophisticated interaction between fact description and charge definitions, such as **MemNet** and **GA_Reader**.
- Methods that augment fact representation with charge definitions through end-to-end schema (**GA_Reader**, **MemNet** and **Ours**) attain better results than **Fact-Law AN**. In addition, compared with **GA_Reader** and **MemNet**, which perform either sentence- or word-level interaction, our approach achieves better performance through considering sentence- and word-level interaction jointly.

Table 2. The experimental results [%] of baselines and our model on two datasets. Four different types of models are separated by lines and the best scores are highlight in bold font. The results are averaged over 5 runs.

Datasets		CAIL150k				CAIL30k			
Model		Acc.	MP	MR	MF1	Acc.	MP	MR	MF1
i	**CNN_classify**	79.23	70.80	62.27	64.97	52.75	23.64	21.95	20.59
	GRU_classify	77.33	72.45	57.42	61.54	56.14	23.99	22.81	21.51
	BERT	77.83	75.43	65.29	67.45	57.92	32.29	30.11	30.25
ii	**Siamese CNN**	72.98	74.52	64.64	66.55	50.66	32.74	33.74	29.28
iv	**TopJudge** [27]	78.56	78.92	58.46	65.32	25.26	25.78	24.32	25.55
	LegalAtt [2]	70.30	76.43	59.48	65.08	51.55	39.81	24.34	26.92
iii	**Fact-Law AN** [16]	75.61	58.89	52.30	53.62	60.73	28.15	25.16	24.79
	GA_Reader	73.78	74.68	66.59	68.21	54.95	39.29	34.05	33.03
	MemNet	80.18	80.09	67.13	70.78	62.40	32.62	27.54	27.64
	Ours	**81.05**	**82.06**	**68.33**	**72.43**	**67.99**	**46.13**	**36.00**	**37.62**

- Our proposed model outperforms other baselines on two datasets. The improvement is especially significant on the `CAIL30k` dataset: our method surpasses the second best about 4.5% in MF1. Since the `CAIL30k` contains more classes with few training samples, the excellent performance of our approach suggests that our auxiliary representations may help to improve the generalization performance for classes with few samples.
- Existing legal models **TopJudge** and **LegalAtt** introduce multiple related tasks and articles for representation training. Although they can improve the performance of charge prediction, **Ours** using charge definitions to relieve the intra-class variance achieves superior performance.

4.5 Ablation Test

We consider several variations of our approach by removing some components of our model to verify the effectiveness of various components in our method. The result is shown in Table 3. As seen, only using fact descriptions without any level auxiliary fact representations (*w/o Hs,Hw*) yields the worst performance, which proves the importance of the use of charge definitions. After adding either the sentence-level (*w/o Hw*) or the word-level auxiliary fact representation (*w/o Hs*), the performance can be significantly improved. We also created a variant of our method without using attention weight g_i of each charge in Eq. (12) in the process of generating charge-token-related fact representation (*w/o Hs,*g_i), which is implemented by setting the attention weight g_i to $\frac{1}{C}$ instead of generated from charge identification part. It can be observed that the performance of *w/o Hs,*g_i declines. This suggests that the two-level interaction is necessary and using them jointly can get the best performance. The little difference between *Ours* and the auxiliary representation only *w/o Ho* shows the importance of original fact representation since it contains original information about the fact description.

Table 3. The experimental results of ablation test of our model on CAIL150k dataset.

Models	Acc.	MP	MR	MF1
Ours	**81.05**	**82.06**	**68.33**	**72.43**
w/o Hs,Hw	77.33	72.45	57.42	61.54
w/o Hw	79.50	78.86	66.18	69.86
w/o Hs	80.62	80.54	66.97	71.28
w/o Hs,g_i	80.54	76.90	64.34	67.98
w/o Ho	80.31	79.12	66.88	70.55

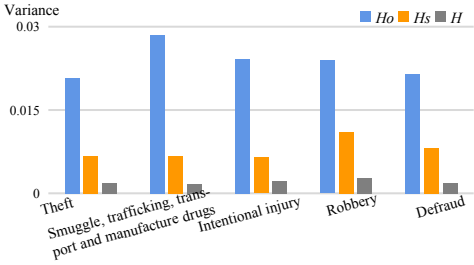

Fig. 3. Intra-class variance of different fact representations of the top-5 frequent classes in CAIL150k dataset. Ho is fact representation only learned from fact description, Hs is the Ho augmented with charge-related fact representation, and H is the Ho augmented with all auxiliary fact representations.

4.6 Intra-class Variance of Different Fact Representations

To investigate whether the fact representation of our method is more stable, we conduct the following experiment: we calculate the variance along each dimension of fact representations from five classes with the most amount of samples, and then use the average variance along all dimensions as an indicator of the intra-class variance of different fact representations. As shown in Fig. 3, fact representation (Ho) only learned from fact description yields the largest intra-class variance. After augmenting fact representation from charge definitions through sentence-level attention (Hs), the intra-class variance declines greatly. Specially, the final fact representation (H) with two auxiliary representations incorporated attains an even lower intra-class variance.

4.7 Case Study

Finally, we select a representative robbery case to give an intuitive illustration of the attention results on the sentence- and word-level interaction. As shown in Fig. 4(a), the case describes that the defendant is convicted of robbery due to stealing property and poking the victim to resist arrest. On the sentence-level interaction, with the increasing of iteration in Eq. (7), our model narrows

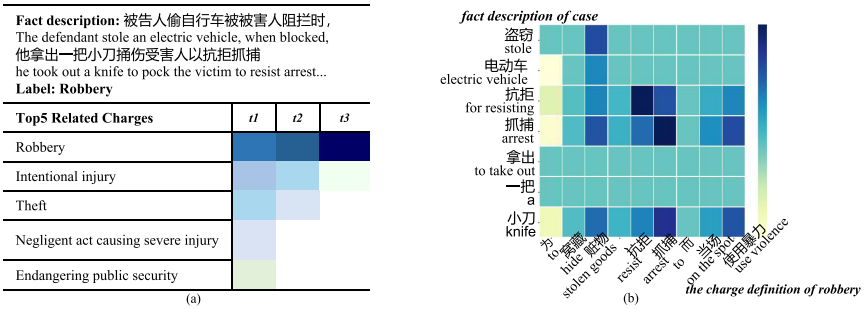

Fig. 4. Attention results of our method for a robbery charge prediction in CAIL150K (in Chinese). The left figure (a) is attention map of sentence-level interaction between fact and charge definitions. t1, t2, and t3 represent the iteration times in Eq. (7). The color darker means the charges are more related to the fact. The right figure (b) is attention map of word-level interaction between fact description and the robbery charge definition. The dark color means a large value.

down the candidate charges and finally identifies the correct related charges. We choose the iteration times as 3 since the performance cannot improve with more iterations. On the word-level interaction, the attention mechanism makes the words in fact description align with the formal terms in charge definitions. Figure 4(b) shows for the words in fact description, which terms are focused on in the charge definition of robbery. The identified keywords in fact description are "electric vehicle", "resisting arrest" and "a knife", which correspond to key terms in robbery definition–"stolen goods", "resist arrest" and "use violence".

5 Conclusion

In this work, we focus on the task of multi-label charge prediction for given fact descriptions of cases. To address the problem of having a large expression variance in fact descriptions due to informal language use, we introduce charge definitions to create auxiliary representations of the fact descriptions by proposed label definitions augmented interaction model. The experimental results on two datasets show the effectiveness of our model on charge prediction. The significant improvement on the dataset with few training data validate that our method can benefit the small sample training scenario and the two-level auxiliary fact representations can help the model to generalize to the unseen description.

Acknowledgments. This work was supported by National Key R&D Program of China (2018YFC0831302), National Natural Sciences Foundation of China (61972386), and Youth Innovation Promotion Association at Chinese Academy of Sciences.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)

2. Bao, Q., Zan, H., Gong, P., Chen, J., Xiao, Y.: Charge prediction with legal attention. In: Tang, J., Kan, M.-Y., Zhao, D., Li, S., Zan, H. (eds.) NLPCC 2019. LNCS (LNAI), vol. 11838, pp. 447–458. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32233-5_35
3. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M., Xu, K.: Adaptive recursive neural network for target-dependent twitter sentiment classification. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (vol. 2: Short papers). vol. 2, pp. 49–54 (2014)
6. Ebesu, T., Shen, B., Fang, Y.: Collaborative memory network for recommendation systems. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 515–524. ACM (2018)
7. Gao, T., Han, X., Liu, Z., Sun, M.: Hybrid attention-based prototypical networks for noisy few-shot relation classification (2019)
8. Hu, Z., Li, X., Tu, C., Liu, Z., Sun, M.: Few-shot charge prediction with discriminative legal attributes. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 487–498 (2018)
9. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. In: ICML Deep Learning Workshop, vol. 2 (2015)
12. Kumar, A., et al.: Ask me anything: Dynamic memory networks for natural language processing. In: International Conference on Machine Learning, pp. 1378–1387 (2016)
13. Lin, W.C., Kuo, T.T., Chang, T.J., Yen, C.A., Chen, C.J., Lin, S.D.: Exploiting machine learning models for Chinese legal documents labeling, case classification, and sentencing prediction. In: Processdings of ROCLING, p. 140 (2012)
14. Liu, C.-L., Hsieh, C.-D.: Exploring phrase-based classification of judicial documents for criminal charges in Chinese. In: Esposito, F., Raś, Z.W., Malerba, D., Semeraro, G. (eds.) ISMIS 2006. LNCS (LNAI), vol. 4203, pp. 681–690. Springer, Heidelberg (2006). https://doi.org/10.1007/11875604_75
15. Liu, C.-L., Liao, T.-M.: Classifying criminal charges in Chinese for web-based legal services. In: Zhang, Y., Tanaka, K., Yu, J.X., Wang, S., Li, M. (eds.) APWeb 2005. LNCS, vol. 3399, pp. 64–75. Springer, Heidelberg (2005). https://doi.org/10.1007/978-3-540-31849-1_8
16. Luo, B., Feng, Y., Xu, J., Zhang, X., Zhao, D.: Learning to predict charges for criminal cases with legal basis. arXiv preprint arXiv:1707.09168 (2017)
17. Sinha, K., Dong, Y., Cheung, J.C.K., Ruths, D.: A hierarchical neural attention-based text classifier. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 817–823 (2018)
18. Sulea, O.M., Zampieri, M., Malmasi, S., Vela, M., Dinu, L.P., van Genabith, J.: Exploring the use of text classification in the legal domain. arXiv preprint arXiv:1710.09306 (2017)
19. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)

20. Wang, S., Mazumder, S., Liu, B., Zhou, M., Chang, Y.: Target-sensitive memory networks for aspect sentiment classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers), pp. 957–967 (2018)
21. Wang, W., Yang, N., Wei, F., Chang, B., Zhou, M.: Gated self-matching networks for reading comprehension and question answering. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers), pp. 189–198 (2017)
22. Weston, J., Chopra, S., Bordes, A.: Memory networks. arXiv preprint arXiv:1410.3916 (2014)
23. Xiao, C., et al.: Cail 2018: A large-scale legal dataset for judgment prediction. arXiv preprint arXiv:1807.02478 (2018)
24. Xiong, C., Merity, S., Socher, R.: Dynamic memory networks for visual and textual question answering. In: International Conference on Machine Learning, pp. 2397–2406 (2016)
25. Yang, W., Jia, W., Zhou, X., Luo, Y.: Legal judgment prediction via multi-perspective bi-feedback network. arXiv preprint arXiv:1905.03969 (2019)
26. Ye, H., Jiang, X., Luo, Z., Chao, W.: Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. arXiv preprint arXiv:1802.08504 (2018)
27. Zhong, H., Zhipeng, G., Tu, C., Xiao, C., Liu, Z., Sun, M.: Legal judgment prediction via topological learning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3540–3549 (2018)