Label Embedding Enhanced Multi-label Sequence Generation Model

Yaqiang Wang[1,2](✉), Feifei Yan[1], Xiaofeng Wang[1], Wang Tang[3,4], and Hongping Shu[1,2](✉)

[1] College of Software Engineering,
Chengdu University of Information Technology, Chengdu 610225, Sichuan, China
{yaqwang,cqshp}@cuit.edu.cn

[2] Sichuan Key Laboratory of Software Automatic Generation and Intelligent Service,
Chengdu 610225, Sichuan, China

[3] School of Electronic Engineering,
Chengdu University of Information Technology, Chengdu 610225, Sichuan, China

[4] Sunsheen Inc., Chengdu 610225, Sichuan, China

Abstract. Existing sequence generation models ignore the exposure bias problem when they apply to the multi-label classification task. To solve this issue, in this paper, we proposed a novel model, which disguises the label prediction probability distribution as label embedding and incorporate each label embedding from previous step into the current step's LSTM decoding process. It allows the current step can make a better prediction based on the overall output of the previous prediction, rather than simply based on a local optimum output. In addition, we proposed a scheduled sampling-based learning algorithm for this model. The learning algorithm effectively and appropriately incorporates the label embedding into the process of label generation procedure. Through comparing with three classical methods and four SOTA methods for the multi-label classification task, the results demonstrated that our proposed method obtained the highest F1-Score (reaching 0.794 on a chemical exposure assessment task and reaching 0.615 on a clinical syndrome differentiation task of traditional Chinese medicine).

Keywords: Multi-label classification · Sequence generation model · Label embedding · Exposure bias problem

1 Introduction

Multi-label classification studies the problem where one real-world object might have multiple semantic meanings by assigning a set of labels to the object in order to explicitly represent its semantics. Multi-label classification has a wide range of real-world application scenarios, and the labels of one object often have correlations. For example, a medical paper often has a set of correlated keywords, which summarizes the topics of the paper's content [1]; a traditional Chinese medicine (TCM) practitioner often uses

Y. Wang and F. Yan—These authors contributed equally to this work.

X. Zhu et al. (Eds.): NLPCC 2020, LNAI 12431, pp. 219–230, 2020.
https://doi.org/10.1007/978-3-030-60457-8_18

multiple correlated syndromes to summarize the chief complaint in a clinical record of TCM for one patient [2].

The multi-label classification task is usually solved by two types of methods. One type is the problem transformation methods, such as the Label Powerset (LP) [3], the Classifier Chain (CC) [4], and another type is the algorithm adaptation methods, such as the ML-kNN [5], the Collective Multi-Label Classifier [6]. In recent years, deep learning has shown excellent performance in various applications, including the multi-label classification task. Researchers attempt to convert the multi-label classification task into a multi-label sequence generation problem through applying the encoder-decoder framework. This approach has yielded satisfactory results [7–9].

The exposure bias problem is often raised when the encoder-decoder framework is applied to the sequence generation task [10]. However, it is ignored when researchers build the multi-label sequence generation models. In consequence, we proposed a novel model in this paper to solve this issue. The model disguises the label prediction probability distribution as label embedding and incorporates each label embedding from previous step into the current step's LSTM decoder process. Furthermore, we proposed a scheduled sampling-based learning algorithm for this model. The experimental results demonstrate that our method outperforms three classical methods, including Binary Relevance (BR), LP and CC, and four SOTA methods, including TextCNN, RCNN, Transformer and SGM, on two representative datasets of the multi-label classification task.

2 Related Work

Considering the label correlation during designing multi-label classification models has attracted much attention. Some work is done by introducing prior knowledge, e.g. the hierarchical relationship among labels [11–14]. Others are done by mining and utilizing the correlations of labels during model training procedure [15–17]. Inspired by the researches of deep learning for machine translation and text summarization, Jinseok et al. [18] proposed to treat the multi-label classification task as a multi-label sequence generation problem and attempted it by using recurrent neural networks. Recently, multi-label sequence generation models based on the encoder-decoder framework have been proposed. Jonas et al. [7] believed that conventional word-level attention mechanism could not provide enough information for the label prediction making, therefore they proposed a multiple attention mechanism to enhance the feature representation capability of input sequences. Li et al. [8] proposed a Label Distributed sequence-to-sequence model with a novel loss function to solve the problem of making a strong assumption on the labels' order. Yang et al. [9] further reduced the sensitivity of the sequence-to-sequence model to the pre-defined label order by introducing reward feedback strategy of reinforcement learning into the model training procedure. However, the exposure bias problem has not been considered, although it is a common issue when the encoder-decoder framework is used to solve the sequence generation problem.

The exposure bias problem is caused by an inconsistency in the training and the inference procedures of the sequence generation models based on the encoder-decoder framework. The inconsistency is reflected in the difference between the input of the next time-step's encoding process in the training procedure and in the inference procedure. One is from the data distribution, and another is from the model distribution. Consequently, when the sequence generation models are applied to the multi-label classification task, the inconsistency would in turn lead to error accumulation during the inference procedure. There are some studies trying to solve the exposure bias problem. Bengio et al. [10] proposed a scheduled sampling algorithm to choose an input for the next time-step from the ground truth word and the predicted word according to a probability change during the sequence generation process. Sam et al. [19] attempted to solve the exposure bias problem through improving the beam search algorithm. Zhang et al. [20] addressed the exposure bias problem by randomly selecting the ground truth word and the predicted word of the previous time-step. An important idea for solving the exposure bias problem is to introduce the predicted words instead of the ground truth words in the training procedure to improve the robustness of the model. How to introduce the predicted words, i.e. the predicted labels, effectively for the multi-label sequence generation models is still an open question.

3 Our Proposed Model

Formally, the multi-label classification task is to assign a label subset $\boldsymbol{y}$, which contains n labels from the label set $\mathcal{L} = \{l_1, l_2, \ldots, l_L\}$, to a sequence $\boldsymbol{x} = \{x_1, x_2, \ldots, x_m\}$, where x_i is the ith word in $\boldsymbol{x}$. From the perspective of a sequence generation model, this multi-label-label classification task can be modeled as finding an optimal label sequence $\boldsymbol{y}^*$ which can maximize the conditional probability:

$$p(\boldsymbol{y}|\boldsymbol{x}) = \prod\nolimits_{t=1}^{n} p(y_t|y_{<t}, \boldsymbol{x}) \tag{1}$$

We apply a sequence-to-sequence model with the attention mechanism for the multi-label sequence generation task. The model in this paper consists of three components, including the XLNet encoder, the attention mechanism and the LSTM decoder. The framework of the model is shown in Fig. 1. $\boldsymbol{h}$, $\boldsymbol{c}$ and $\boldsymbol{s}$ in Fig. 1 represent the hidden states of the encoder, the context vector, and the hidden states of the decoder, respectively, and the subscript t in the figure represent the time-step.

3.1 The XLNet Encoder

Different from Jonas et al. [7], we apply the generalized autoregressive language model, XLNet [21], to replace the commonly used Bidirectional LSTM and GRU encoders in this paper. The XLNet will output the hidden state vector $\boldsymbol{h}_i$ for each word.

$$\boldsymbol{h}_i = \mathrm{XLNet}(x_i) \tag{2}$$

3.2 The Attention Mechanism

Different words in one sequence often have different contributions when the model predicts the labels. The attention mechanism can make the model have ability to give

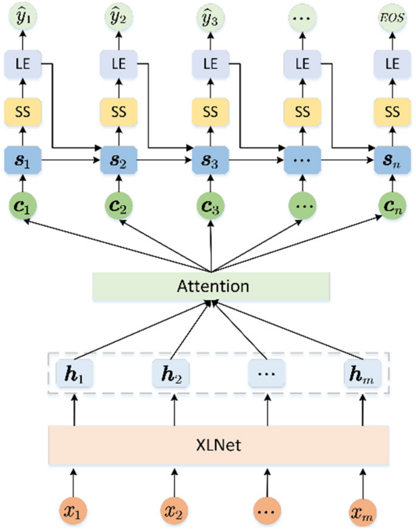


Fig. 1. Framework of our proposed model. LE denotes the label embedding method and SS denotes the scheduled sampling process.

different weights to different words of a sequence according to the contributions of the words to the label prediction task. The weight α_{ti} of a word x_i in a sequence $\boldsymbol{x}$ at time-step t is calculated by

$$\alpha_{ti} = \boldsymbol{v}^T \tanh(\boldsymbol{W}_1 \boldsymbol{s}_{t-1} + \boldsymbol{U}_1 \boldsymbol{h}_i), \tag{3}$$

where $\boldsymbol{s}_{t-1}$ is the hidden state of the decoder at time-step $t - 1$ and $\boldsymbol{v}^T$, $\boldsymbol{W}_1$ and $\boldsymbol{U}_1$ are the weighting parameters. The weights will be normalized by using the SoftMax function

$$w_{ti} = \frac{\exp(\alpha_{ti})}{\sum_{j=1}^{m} \exp(\alpha_{tj})}, \tag{4}$$

and then the final context vector $\boldsymbol{c}_t$ is computed as follows:

$$\boldsymbol{c}_t = \sum\nolimits_{i=1}^{m} w_{ti} \boldsymbol{h}_i \tag{5}$$

3.3 The LSTM Decoder

LSTM models the correlations between labels at different time-steps in the generated label sequence. The context vector $\boldsymbol{c}_t$, the hidden state $\boldsymbol{s}_{t-1}$ of the decoder at time-step t − 1 and the label embedding, which will be introduced in Sect. 4, form the input to the hidden state $\boldsymbol{s}_t$ of the decoder at time-step t as follows

$$\boldsymbol{s}_t = \text{LSTM}\left(\boldsymbol{s}_{t-1}, \left[\boldsymbol{c}_t; g\left(P_{t-1}^{y}\right)\right]\right), \tag{6}$$

where P_{t-1}^{y} represents the label prediction probability distribution for the labels outputted by the LSTM decoder at time-step $t - 1$, [;] is the vector concatenation operation, and $g(\cdot)$ is used to disguise the label prediction probability distribution as a label embedding, which will be introduced in the next section.

4 Label Embedding Method

Inspired by the Global Embedding [22] and the LSTM gating mechanism [23], we proposed a label embedding method which is used to disguises the label prediction probability distribution of the labels outputted by the LSTM decoder at time-step $t-1$. The label embedding outputted from $g\left(P_t^y\right)$ is formed by an expected label embedding $\bar{\boldsymbol{e}}_t$ at time-step t and a label embedding $\hat{\boldsymbol{e}}_t$ of which label with the highest probability in P_t^y.

$$g\left(P_t^y\right) = \left[\boldsymbol{o}_t \odot \bar{\boldsymbol{e}}_t; (1-\boldsymbol{o}_t) \odot \hat{\boldsymbol{e}}_t\right] \tag{7}$$

$$\bar{e}_t = P_t^y \boldsymbol{E}, \tag{8}$$

$$P_t^y = \mathrm{SoftMax}\left(\frac{\boldsymbol{s}_{t-1}\boldsymbol{W}_2}{\gamma}\right) \tag{9}$$

$$\boldsymbol{o}_t = \sigma\left(\boldsymbol{W}_3\bar{\boldsymbol{e}}_t + \boldsymbol{W}_4\hat{\boldsymbol{e}}_t\right) \tag{10}$$

where $\odot$ is the element-wise multiplication operation, $\hat{\boldsymbol{e}}_t$ is selected from $\boldsymbol{E} \in \mathbb{R}^{k\times L}$, which is a learnable embedding matrix, k is the dimension of the label embeddings, $\boldsymbol{W}_2 \in \mathbb{R}^{d\times L}$ is a weight matrix, d is the dimension of the hidden state of the LSTM decoders. The large L is, the more elements in P_t^y tend to be zero. It would, consequently, causes the back-propagation process having the vanishing gradient problem. This is why we define P_t^y in terms of Eq. (9), and the Eq. (9) is inspired by the Scaled Dot-Product Attention method [24], where γ is a scaling factor used to solve the aforementioned problem. $\sigma(\cdot)$ is the sigmoid function, and $\boldsymbol{W}_3$, $\boldsymbol{W}_4 \in \mathbb{R}^{k\times k}$. The range of the values of $\boldsymbol{o}_t$ are in (0, 1). $\boldsymbol{o}_t$ and $(1-\boldsymbol{o}_t)$ define the contributions of $\bar{\boldsymbol{e}}_t$ and $\hat{\boldsymbol{e}}_t$, and $\boldsymbol{o}_t$ will be automatically determined by the learning algorithm.

5 Learning Algorithm

In this section, we designed the learning algorithm for the proposed model based on a scheduled sampling process. The cross-entropy loss function is used in this paper, and it is defined as follows:

$$loss_{CE} = -\sum\nolimits_{t=1}^{n} \log p_\theta(y_t|y_{t-1};\boldsymbol{x}), \tag{11}$$

$$\hat{y}_t = \mathrm{argmax}_y p_\theta\left(y|\hat{y}_{t-1}\right), \tag{12}$$

where θ is the set of parameters to be learned, $\hat{y}_t$ represents the predicted label at time-step t. In order to learn the parameters based on variable length sequences, following the method used in [22], we also added a special token, <EOS> , at the end of each sequence.

The scheduled sampling approach has been proven to be effective for solving exposure bias problem [10]. Therefore, we followed this idea and proposed a scheduled sampling-based algorithm for our proposed multi-label sequence generation model. The pseudo code is described in Algorithm 1.

Algorithm 1 Algorithm Combining Label Embedding and Scheduled Sampling

```
Input: $(\boldsymbol{x}, \boldsymbol{y} = \{y_1, y_2, \cdots, y_n\})$, $threshold$, $k$
Output: $\{\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n\}$
for $i$ in epoch do
    $t \leftarrow 1$
    if $i \leq threshold$ then
        $\hat{e}_t = y_t$
        $g(P_t^y)$
    else
        $\epsilon_i = k^{(i-threshold)}$
        $\hat{e}_t = ScheduledSampling(\epsilon_i, \hat{y}_t, y_t)$
        $g(P_t^y)$
    end if
    $t \leftarrow t + 1$
end for
```

If the label embedding method introduced in Sect. 4 is utilized in the early stages of the training procedure, it may bring too much uncertainty to the loss leading to loss fluctuation and may even cause the curve of the loss function to not converge. Therefore, we designed a function of the number of the iteration index *i*, $\in_i = k^{(i-threshold)}$, which is used to control that only the ground labels will be used in the early stages of the training procedure, and after a period of training time, the label embedding will be incorporated. In $\in_i$, *k* is a hyperparameter which is ranging from 0 to 1, and *threshold* is the number of iterations that the algorithm starts using the scheduled sampling algorithm to get $\hat{\boldsymbol{e}}_t$. It is clear that the value of $\in_i$ begins to decay exponentially after the number of iterations reaching the *threshold*.

6 Experiments

In this paper, we compared our proposed method with three classical multi-label classification methods and four SOTA methods on two biomedical domain datasets. One is in Chinese, and another is in English. The datasets, the evaluation measurements, the compared methods and the results will be introduced in following sections.

6.1 Datasets

We used two biomedical domain datasets in the experiments. Both of the datasets are typically used to validate the multi-label classification methods. Detailed information of these datasets is shown in Table 1. CEA (a Chemical Exposure Assessments dataset) is an English dataset, and TCM (a syndrome differentiation dataset of traditional Chinese medicine) is a Chinese dataset.

CEA: PubMed [28] provides a large amount of biochemical exposure information, which is of vital research value for the study of human health. Larsson et al. [25] constructed the CEA dataset relying on the domain experts based on part of PubMed literature. The CEA dataset contains 32 labels which are keywords described from the perspectives of biological detection and exposure pathway.

Table 1. Detailed statistics information of the datasets CEA and TCM.

Dataset	Number of labels	Number of instances	Number of words in one instance			Number of labels in one instance		
			Avg	Max	Min	Avg	Max	Min
CEA	32	3661	233.6	622	49	2.0	8	0
TCM	1127	10000	8.84	35	1	1.85	5	1

TCM: The TCM dataset is composed of chief complaints and syndromes. The chief complaints are noted by TCM experts during their daily work, and they are short and concise texts. The syndromes are descriptive and positional order sensitive, and they are the labels. The dataset is obtained from a real-world medical information system. An example is list as follows:

A chief complaint: “心悸, 胸闷, 气短,口干, 不渴, 左胁略胀, 饮食正常，二便正, 常, 舌暗红, 形体胖 苔薄, 脉时快时慢, 节律不齐”. (Palpitation, chest distress, breathe hard, dry mouth, hydroadipsia, left rib-side distention, normal diet, bowel function is normal, dark red and swollen tongue, thin tongue fur, pulse waxes and wanes, rhythm not neat).

Syndrome labels: “痰热内扰, 心气不足”. (Phlegm hot inside, heart qi insufficient).

6.2 Evaluation Measurements

There are two types of evaluation measurements for the multi-label classification task. They are sample-based measure and label-based measure. In this paper, we used the label-based measurements, including Precision$_{micro}$ (P$_{micro}$), Recall$_{micro}$ (R$_{micro}$), and F1$_{micro}$, to evaluate the performance of different methods. The calculating methods of P$_{micro}$, R$_{micro}$ and F1$_{micro}$ are shown in Eq. (13), (14) and (15), respectively.

$$\mathrm{P}_{micro} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{13}$$

$$\mathrm{R}_{micro} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{14}$$

$$\mathrm{F1}_{micro} = \frac{2 \times \mathrm{P}_{micro} \times \mathrm{R}_{micro}}{\mathrm{P}_{micro} + \mathrm{R}_{micro}} \tag{15}$$

6.3 Experimental Settings

CEA and TCM datasets are randomly divided into three parts, including a training dataset, a validation dataset and a test dataset, with a ratio of 7:1:2. The learning rate of XLNet is set to 3e−5, the learning rate of other layers in the model is set to 0.001, we used the Adam optimizer, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The batch size is set to 16, the hyperparameter k in the learning algorithm is set to 0.85, and the dropout and L2 regularizer are used to avoid overfitting. The dimension of pre-trained XLNet word embedding is 768.

Three classical multi-label classification models, i.e. BR, LP and CC, are implemented by using Scikit-Multilearn [26], and LinearSVM is used in these models as the base classifier. The descending order of label's frequencies is used in CC. TextCNN and RCNN are implemented based on an open source tool, named NeuralNLP [27]. We used the SGM code published by Yang et al. [21] in this paper, and the default parameter setting, which can yield the best result, is used.

6.4 Results

The best $\mathrm{F1}_{micro}$ results achieved by different methods under different settings are listed in Table 2.

Table 2. Comparison of different results of various methods.

Algorithms	CEA			TCM		
	P_{micro}	R_{micro}	$\mathrm{F1}_{micro}$	P_{micro}	R_{micro}	$\mathrm{F1}_{micro}$
BR	0.896	0.555	0.685	**0.843**	0.402	0.544
CC	**0.897**	0.547	0.679	0.764	0.460	0.574
LP	0.669	0.483	0.561	0.606	0.609	0.608
TextCNN	0.740	0.643	0.688	0.800	0.487	0.605
RCNN	0.757	0.669	0.710	0.667	0.489	0.564
Transformer	0.629	0.590	0.609	0.713	0.484	0.576
SGM	0.590	0.584	0.586	0.559	0.566	0.552
SGM+XLNet	0.792	0.781	0.787	0.588	0.600	0.594
Our	0.796	0.776	0.786	0.610	0.597	0.604
+SS	0.788	**0.788**	0.788	0.614	0.603	0.608
+LE	0.801	0.776	0.789	0.628	0.593	0.610
+LE+SS	0.813	0.777	**0.794**	0.620	**0.611**	**0.615**

"Our" represents our proposed method, LE = Label Embedding, SS = Scheduled Sampling

In general, it vividly shows in Table 2 that the proposed method outperforms other methods. On the CEA dataset, the best $\mathrm{F1}_{micro}$ (Our+LE+SS) obtained by our method can reach 0.794, which is 0.149 higher than other methods on average. On the TCM dataset, the best $\mathrm{F1}_{micro}$ (Our+LE+SS) can reach 0.615, which is also higher than other methods, but is a little bit lower than on the CEA dataset, it is because the label set size of the TCM dataset is much larger than the CEA dataset.

The P_{micro} and R_{micro} results of SGM and our proposed method listed in Table 2 show that converting the multi-label classification tasks into a multi-label sequence generation problem can achieve more balanced P_{micro} and R_{micro} results. Almost all other methods have the problem of high P_{micro} and low R_{micro}.

Compared with SGM, our proposed method is much better. On one hand, XLNet used in our method has a stronger encoding capacity than bidirectional LSTM used in SGM, and XLNet can achieve good results with only limited sample fine-tuning. On the other hand, our proposed label embedding method and the scheduled sampling-based learning algorithm further improve the performance.

Through a further in-depth analysis of the results, we found that the unseen domain-specific terms are a potential negative factor for the performance improvement. Taking the results on the TCM dataset as an example, the dataset contains a large number of domain-specific terms, e.g. "脉细" (pulse fine), "神疲" (mental fatigue), etc., which are usually unseen in the vocabulary used in XLNet, because the XLNet is pre-trained on a general domain corpus. Consequently, it would result in many inaccurate semantic representations for these domain-specific terms and lead to a negative impact on the performance.

Comparison of the Label Generation Results with Different Granularity. The labels (i.e. the syndromes) in the TCM dataset are often composed of fine-grained semantic units (characters or syndrome factors), e.g. syndrome factors "筋脉" (tendons), "瘀" (stasis) and "滞" (stagnation) making up the syndrome "筋脉瘀滞" (tendons stasis and stagnation). Therefore, we attempt to reduce the label set size by fine-grained labels. With this approach, we expect to further improve the performance. The results are listed in Table 3.

Table 3. Comparison of different granularity label generation results on the TCM dataset.

Granularity	P_{micro}	R_{micro}	$\mathrm{F1}_{micro}$
Character level	0.513	0.322	0.396
Syndrome factor level	0.572	0.479	0.522

The results in Table 3 are worse than in Table 2. Character level's results are worse than syndrome factor level's results, and both are worse than syndrome level's results. This result is mainly due to the fact that our proposed method cannot accurately generate complete syndrome labels based on the fine-grained labels. This issue presents us a new challenge that how to generate coarse-grained labels accurately based on the fine-grained labels. This is our future work.

Comparison of the Loss Convergence Results. In order to verify the contribution of the label embedding, we further examined the loss convergence results during training procedure (shown in Fig. 2).

In Fig. 2, TCM and CEA represent that label embedding and scheduled sampling are used during training. TCM+LE (truth) and CEA+LE (truth) represent just use the label embedding of the ground truth from previous time-step. TCM+LE (predict) and

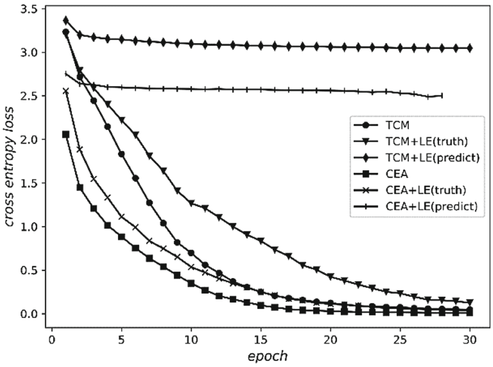
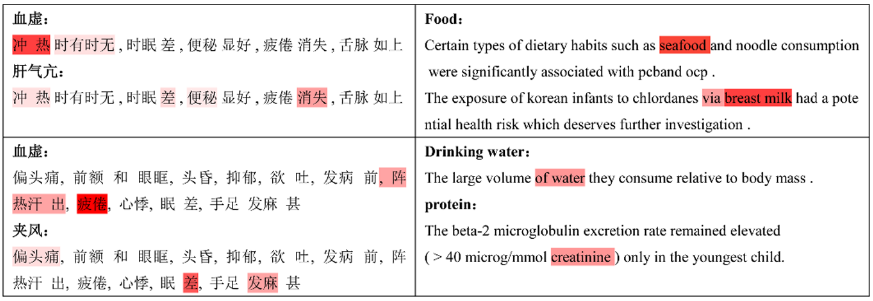


Fig. 2. Comparison of the loss convergence results with and without the label embedding.

CEA+LE (predict) represent just use the label embedding of the prediction with highest probability from previous time-step. It is clear that our proposed method has better loss convergence results on both datasets.

Qualitative Evaluation Results of Attention in Our Method. Figure 3 visualizes four examples of the attention results. The left is examples of TCM, and the right is examples of CEA. The results show that the attention mechanism is able to accurately make use of corresponding key informative words in the sequence when predicting labels.



血虚: 冲 热 时有时无 , 时眠 差 , 便秘 显好 , 疲倦 消失 , 舌脉 如上 **肝气亢:** 冲 热 时有时无 , 时眠 差 , 便秘 显好 , 疲倦 消失 , 舌脉 如上	**Food:** Certain types of dietary habits such as seafood and noodle consumption were significantly associated with pcband ocp . The exposure of korean infants to chlordanes via breast milk had a potential health risk which deserves further investigation .
血虚: 偏头痛, 前额 和 眼眶, 头昏, 抑郁, 欲 吐, 发病 前, 阵热汗 出, 疲倦, 心悸, 眠 差, 手足 发麻 甚 **夹风:** 偏头痛, 前额 和 眼眶, 头昏, 抑郁, 欲 吐, 发病 前, 阵热汗 出, 疲倦, 心悸, 眠 差, 手足 发麻 甚	**Drinking water:** The large volume of water they consume relative to body mass . **protein:** The beta-2 microglobulin excretion rate remained elevated (> 40 microg/mmol creatinine) only in the youngest child.

Fig. 3. Four examples of the attention results visualized based on heatmaps.

7 Conclusion

Multi-label classification has a wide range of real-world application scenario. It is an effective way to treat the multi-label classification as a multi-label sequence generation task, and it is of great significance to use other auxiliary information (such as the label embedding) to enhance the ability of multi-label sequence generation. The experimental results show that our proposed label embedding method and the scheduled sampling-based learning algorithm are effective and outperform the compared method.

Acknowledgement. The research work is partially supported by the Sichuan Major Science and Technology Special Program under Grant (2017GZDZX0002), the Sichuan Science and Technology Program under Grant (2018GZ207), the Sichuan Province Science and Technology Support Program under Grant (2020YFG0299, 2020YFSY0067), and the National Natural Science Foundation of China under Grant (61801058, 61501063).

References

1. Pratt, W., Yetisgen-Yildiz, M.: LitLinker: capturing connections across the biomedical literature. In: 2nd International Conference on Knowledge Capture, New York, pp. 105–12. Association for Computing Machinery (2003)
2. Zhang, L., Yu, D.L., Wang, Y.G.: Selecting an appropriate interestingness measure to evaluate the correlation between Chinese medicine syndrome elements and symptoms. Chin. J. Integr. Med. **18**(2), 93–99 (2012)
3. Tsoumakas, G., Vlahavas, I.: Random *k*-labelsets: an ensemble method for multilabel classification. In: Kok, J.N., Koronacki, J., de Mantaras, R.L., Matwin, S., Mladenič, D., Skowron, A. (eds.) ECML 2007. LNCS (LNAI), vol. 4701, pp. 406–417. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74958-5_38
4. Read, J., Pfahringer, B., Holmes, G., et al.: Classifier chains for multi-label classification. Mach. Learn. **85**(3), 333 (2011)
5. Zhang, M.L., Zhou, Z.H.: ML-KNN: a lazy learning approach to multi-label learning. Pattern Recognit. **40**(7), 2038–2048 (2007)
6. Ghamrawi, N., McCallum, A.: Collective multi-label classification. In: 14th ACM International Conference on Information and Knowledge Management, New York, pp. 95–200. Association for Computing Machinery (2005)
7. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. In: 34th International Conference on Machine Learning, Sydney, pp. 1243–1252 (2017)
8. Li, W., Ren, X.C., Dai, D., et al.: Sememe prediction: learning semantic knowledge from unstructured textual wiki descriptions. arXiv preprint arXiv:1808.05437 (2018)
9. Yang, P.C., Luo, F.L., Ma, S.M., et al.: A deep reinforced sequence-to-set model for multilabel classification. In: 57th Annual Meeting of the Association for Computational Linguistics, Florence, pp. 5252–5258. Association for Computational Linguistics (2019)
10. Bengio, S., Vinyals, O., Jaitly, N., et al.: Scheduled sampling for sequence prediction with recur-rent neural networks. In: Advances in Neural Information Processing Systems, Montreal, pp. 1171–1179. Neural Information Processing Systems (2015)
11. Kowsari, K., Brown, D.E., Heidarysafa, M., et al.: HDLTex: hierarchical deep learning for text classification. In: 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, pp. 364–371. IEEE (2017)
12. Baker, S., Korhonen, A.L.: Initializing neural networks for hierarchical multi-label text classification. In: 16th Biomedical Natural Language Processing Workshop, Vancouver, pp. 307–315. Association for Computational Linguistics (2017)
13. Peng, H., Li, J., He, Y., et al.: Large-scale hierarchical text classification with recursively regularized deep graph-CNN. In: 2018 World Wide Web Conference, Lyon, France pp. 1063–1072. International World Wide Web Conferences Steering Committee (2018)
14. Cerri, R., Barros, R.C., De Carvalho, A.C.: Hierarchical multi-label classification using local neural networks. J. Comput. Syst. Sci. **80**(1), 39–56 (2014)

15. Yang, Y.Y., Lin, Y.A., Chu, H.M., et al.: Deep learning with a rethinking structure for multi-label classification. In: Asian Conference on Machine Learning, Nagoya, pp. 125–140. Proceedings of Machine Learning Research (2019)
16. Fu, D., Zhou, B., Hu, J.: Improving SVM based multi-label classification by using label relationship. In: 2015 International Joint Conference on Neural Networks (IJCNN), Killarney, pp. 1–6. IEEE (2015)
17. Aly, R., Remus, S., Biemann, C.: Hierarchical multi-label classification of text with capsule networks. In: 57th Annual Meeting of the Association for Computational Linguistics, Florence, pp. 323–330. Association for Computational Linguistics (2019)
18. Nam, J., Mencía, E.L., Kim, H.J., et al.: Maximizing subset accuracy with recurrent neural networks in multi-label classification. In: Advances in Neural Information Processing Systems, Long Beach, pp. 5413–5423. Neural Information Processing Systems (2017)
19. Wiseman, S., Rush, A.M.: Sequence-to-sequence learning as beam-search optimization. In: 2016 Conference on Empirical Methods in Natural Language Processing, Austin, pp. 1296–1306 (2016)
20. Zhang, W., Feng, Y., Meng, F., et al.: Bridging the gap between training and inference for neural machine translation. arXiv preprint arXiv:1906.02448 (2019)
21. Yang, Z., Dai, Z., Yang, Y., et al.: XLNet: generalized autoregressive pretraining for language understanding. In: Advances in Neural Information Processing System, Vancouver, pp. 5753–5763. Neural Information Processing Systems (2019)
22. Yang, P., Sun, X., Li, W., et al.: SGM: sequence generation model for multi-label classification. In: 27th International Conference on Computational Linguistics, Santa Fe, pp. 3915–3926. Association for Computational Linguistics (2018)
23. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
24. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, Long Beach, pp. 5998–6008. Neural Information Processing Systems (2017)
25. Larsson, K., Baker, S., Silins, I., et al.: Text mining for improved exposure assessment. PLoS ONE **12**(3), e0173132 (2017)
26. Szymański, P., Kajdanowicz, T.: A scikit-based Python environment for performing multi-label classification. arXiv preprint arXiv:1702.01460 (2017)
27. Liu, L., Mu, F., Li, P., et al.: NeuralClassifier: an open-source neural hierarchical multi-label text classification toolkit. In: 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Florence, pp. 87–92. Association for Computational Linguistics (2019)
28. PubMed Homepage. https://pubmed.ncbi.nlm.nih.gov. Accessed 20 Mar 2020