Imbalanced Chinese Multi-label Text Classification Based on Alternating Attention

Hongliang Bi, Han Hu, Pengyuan Liu*
Beijing Language and Culture University, Beijing 100083, China
Language Resources Monitoring & Research Center, Beijing 100083, China
{201821198617, 201821198609}@stu.blcu.edu.cn, liupengyuan@blcu.edu.cn

Abstract

In this work, we construct an imbalanced Chinese multi-label text classification dataset, IMCM. The imbalance is mainly reflected in: (1) The degree of discrimination among labels is different. (2) The distribution of labels is moderately imbalanced. Then, we adopt several methods for multi-label classification and conduct thorough evaluation of them, which show that even the most competitive models struggle on this dataset. Therefore, to tackle these imbalanced problems, we proposed an alternating attention model, AltXML. Two attention heads which alternately reading sequence enable the model capture different parts of the document rather than one point. Experimental results show that our proposed model significantly outperforms the state-of-the-art baselines in our IMCM dataset, and also achieves quite good results in several public datasets.

1 Introduction

Multi-label classification (MLC) is an important task in natural language processing (NLP) due to the increasing number of fields where it can be applied, such as text classification, tag suggestion, information retrieval, and so on. Compared to single-label classification task, multi-label classification task aims to assign a set of labels to a single instance simultaneously. However, the number of label sets grows exponentially as the number of class labels increases and the uncertainty in the number of labels per instance inevitably makes the MLC task much more difficult to solve. Therefore, the key challenge of this task lies in the overwhelming and uncontrollable size of output space.

Large amount of efforts have been done towards MLC task, including Binary Relevance (BR) (Boutell et al., 2004), Classifier Chains (CC) (Read et al., 2011), Label Powerset (LP) (Tsoumakas and Vlahavas, 2007), PD-Spare (Yen et al., 2016), SLEEC (Bhatia et al., 2015), AnnexML (Tagami, 2017), PfastreXML (Jain et al., 2016), Parabel (Prabhu et al., 2018).. In addition to the above methods, neural networks provide some new approaches: CNN (Kim, 2014), CNN-RNN (Chen et al., 2017), SGM (Yang et al., 2018), etc. These methods have made great progress in capturing label correlations to cope with the exponential-sized output space, but still face the problem of high computational complexity and poor scalability.

While utilizing correlations among labels is essential for MLC task, in real-word scenarios, there are no obvious semantic boundaries among some labels and some seemingly distinct labels may appear together, especially for text. Moreover, the distribution of labels may be imbalanced. On the one hand, the number of instance belonging to a certain label may outnumber other labels. On the other hand, there may be a relatively high number of examples associated with the most common labels or infrequent labels (Gibaja and Ventura, 2015). These may affect the performance of models utilizing correlations of labels. Therefore, it is important to explore the balance between using correlation to reduce output space and improving the ability to refine labels.

* Corresponding Author.

We inspect the commonly used multi-label text classification datasets consist of Rcv1v2 (Lewis et al., 2004), AAPD (Yang et al., 2018), etc. Some of them has been used as benchmarks, but still can not meet the actual demand. The numbers of class labels or labels per instance is small, and the semantic boundaries among the labels are obvious to some extend. Therefore, to further explore this field, we propose an imbalanced Chinese multi-label text classification dataset, IMCM [1].

Furthermore, we conduct a detailed evaluation for diverse MLC models on our dataset and two public datasets. Experimental results show that several models that perform well on other datasets struggle on our dataset. Our point of view is that, different from single label classification models which need to focus on the most important part of the document, multi-label classification models need to be aware of different parts. That means that models can't be bound by a certainly associated label.

Therefore, inspired by the idea of dilated convolution which has become popular in semantic segmentation (Yu and Koltun, 2016), we propose our alternating attention model, AltXML. Two attention heads which alternate reading sequence enable the model capture different parts of the document rather than one point. We evaluate our model on different datasets. Comparison with other models indicates that the trade-off between using correlation to reduce output space and improving the ability to refine labels needs further research. In summary, our contribution is three-fold:

- We construct an imbalanced Chinese multi-label text classification dataset, IMCM.
- We implement diverse MLC models and propose our alternating attention model.
- We conduct a detailed evaluation for these models on three datasets with different imbalance ratios, by comparing on them, our model achieves promising performance.

2 Related Work

Multi-label classification studies the problem where each example is represented by a single instance while associated with a set of labels simultaneously. There are two main types of methods for MLC task: problem transformation methods and algorithm adaptation methods.

Binary Relevance (BR) transforms the task of multi-label classification into the task of binary classification, which is simple and reversible but ignores potential correlations among labels and may lead to the issue of sample imbalance. Label powerset (LP) generates a new class for each possible combination of labels and then solves the problem as a single-label multi-class one. Classifier Chains (CC) treats this task as a sequence labeling problem and overcomes the label independence assumption of BR due to classifiers are built upon the previous predictions. In addition to traditional machine learning methods, Neural networks provide some new approaches to MLC task. These methods have made great progress in multi-label classification task, but still face the problem of high computational complexity and poor scalability to meet high-order label correlations.

CNN uses multiple convolution kernels to extract text feature, which is then input to the linear transformation layer followed by a sigmoid function to output the probability distribution over the label space. CNN-RNN incorporated CNN and RNN so as to capture both global and local semantic information and model high-order label correlations.

Nam et al. (2017) also treat the multi-label classification task as a sequence labeling problem but replace classifier chains with RNN. It allows to focus on the prediction of the positive labels only, a much smaller set than the full set of possible labels. Yang et al. (2018) propose to view the MLC task as a sequence generation problem to take the correlations between labels into account.

Typically, there are two main available multi-label text classification datasets, which all stem from English reading materials. Rcv1v2 (Lewis et al., 2004) is widely used in multi-Label classification task. It consists more than 80,000 manually classified English newswire stories, which divided by Lin et al. (2018). The total number of topic labels is 103.

AAPD (Yang et al., 2018) is a large English multi-label text classification dataset. It contains abstract and corresponding topics of 55,840 papers in the computer science field on the Arxiv. The total number of subjects is 54.

[1] https://github.com/NLPBLCU/imcm-dataset

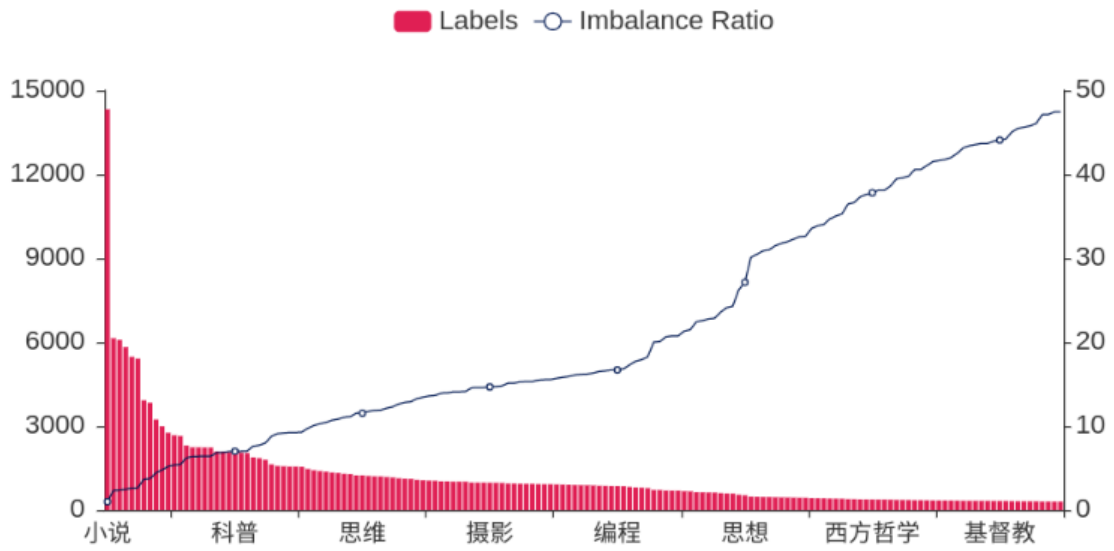

Figure 1: Distribution and Imbalanced Ratio of labels on IMCM dataet. Imbalanced Ratio is the ratio of the frequency of the label to the highest frequency.

Datasets	Inst.	Lab	Card.	Dens.	Len.	IR.	Train Set	Valid Set	Test Set
Rcv1v2	804,414	103	3.24	0.031	123.94	17.44	802,414	1,000	1,000
AAPD	55,840	54	2.4	0.044	163.43	6.58	53,840	1,000	1,000
IMCM	52,052	158	3.7	0.023	348.91	10.35	41,642	5,205	5,205

Table 1: Comparison of IMCM dataset with existing MLC datasets. Inst and Lab denote the total number of instances and labels, respectively. Card means the average number of labels per instance. DENS normalizes Card by the Lab. Len refers to the average length of the instance. IR indicates how imbalanced the top 50 percentage of labels are.

3 IMCM Dataset

For the purpose of constructing highly reliable multi-label text classification dataset, we have collected nearly 60,000 books' information from Douban [2], which consists of content summary and author introduction. Labels of each book are manually marked by members of Douban. Unlike the above described datasets, the difference among some labels in the IMCM is very subtle, such as Humanistic and Human nature. And distribution of labels is very imbalanced, which can be seen in figure 1. These characteristics make it not feasible for labels to be classified in an extensive way. Therefore, we limited the number of words per instance no less than 50 to provide adequate information. Finally, we got 52,286 documents.

In order to evaluate the data effectively, we carry on the same distribution sampling to the data. In the end, we got 41,829 training data, 5,228 validation data and 5,229 test data. The total number of labels is 158, the average number of labels per instance is 3.7 (can be seen in figure 2), the average length of the instance is 348.91 and the imbalanced ratio

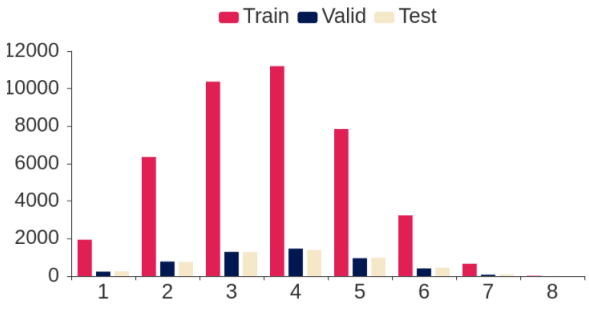

Figure 2: Distribution of the number of labels per instance.

[2] https://book.douban.com

of labels is 10.35. Comparison of IMCM dataset with existing MLC datasets can be seen in Table 1. We can see that our dataset is longer than the other two. Besides, neither like the extreme imbalance of the labels of the Rcv1v2 dataset nor like the small-scale imbalance of the labels of the AAPD dataset, our dataset makes a trade-off. This avoids the overwhelming interference caused by the extreme imbalance of data, and allows us to make some explorations on this basis.

4 Alternating Attention Model

We introduce our proposed model in detail in this section. First, we give an overview of the model in Figure 3. It consists of four layers: Word Representation Layer, Bidirectional LSTM Layer, Alternating Attention Layer, and Classification Layer.

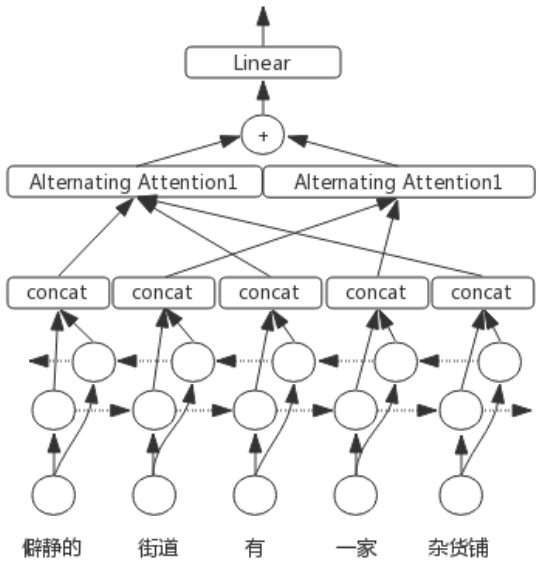

Figure 3: Overview of the AltXML model

4.1 Word Representation Layer

The input of AltXML is raw tokenized text, each word is represented by word embedding. Let T and d respectively represent the length of the input text and the dimension of word representation. The output of word representation as follows:

$$X = (x_1, x_2, ..., x_T)$$

where x_t is a dense vector for each word.

4.2 Bidirectional LSTM Layer

We use a Bidirectional LSTM (Hochreiter and Schmidhuber, 1997) to capture both the left-sides and right-sides context at each time step, the output of BiLSTM can be obtained as follows:

$$\overrightarrow{h_t} = LSTM(x_t, \overrightarrow{h_t}, C_{t-1})$$

$$\overleftarrow{h_t} = LSTM(x_t, \overleftarrow{h_t}, C_{t-1})$$

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$$

where h_t is obtained by concatenating forward $\overrightarrow{h_t}$ and backward $\overleftarrow{h_t}$.

4.3 Alternating Attention Layer

We alternately send the output of the BiLSTM to the two attention layers, reduce the coupling between information, so that it is able to remove the negative effects such as information loss caused by general attention mechanism, such as focus on one key point. The output of alternating attention can be obtained as follows:

$$m_{2i} = \frac{e^{h_{2i} w_m^T}}{\sum_{t=1}^{T} e^{h_{2i} w_m^T}} ; m_{2i+1} = 0$$

$$n_{2i+1} = \frac{e^{h_{2i+1} w_n^T}}{\sum_{t=1}^{T} e^{h_{2i+1} w_n^T}} ; n_{2i} = 0$$

$$a = \sum_{i=1}^{T} Relu(m + n) * h_i$$

where m_i and n_i is the normalized coefficient of h_i.

Besides, it is able to expand the attention at the polynomial level without increasing the number of parameters. Thus, it becomes possible for alternating attention to capture longer-term dependency and avoid gridding effects caused by dilation.

4.4 Classification Layer

AltXML has one fully connected layers as output layer. Then, predicted probability $\hat{y}$ for the label can be obtained as follows:

$$\hat{y} = f(aw^T + b)$$

where, function f is sigmoid activation function.

4.5 Loss Function

We uses the binary cross-entropy loss function, which was used in XML-CNN (Liu et al., 2017) as the loss function. The loss function is given as follow:

$$L(\theta) = -\frac{1}{NL}\sum_{i=1}^{N}\sum_{j=1}^{N} y_{ij}\log(\hat{y}_{ij}) + (1-y_{ij})\log(1-\hat{y}_{ij})$$

where N is the number of samples, L is the number of labels, $\hat{y}_{ij} \in [0, 1]$ and $y_{ij} \in \{0, 1\}$ are the predicted probability and true values, respectively, for the i-th sample and the j-th label.

5 Experiments

5.1 Setting

Training details of neural network models are illustrated as follows.

- Vocabulary: For training efficiency and generalization, in all datasets, we truncate the full vocabulary and set a shortlist of 60,000. Note that, for Chinese, we use Jieba [3] to cut words and not use domain dictionary.
- Embedding layer: We set word embedding dimension to 256 and use randomly initialized embedding matrix with the normal distribution $\mathcal{N}(0, 1)$. Note that, no pre-trained word embeddings are used in our experiments.
- BiLSTM layer: We use single-layered bidirectional LSTM that output dimension in each direction is 100, and randomly initialized it with uniform distribution $\mathcal{U}(-\sqrt{k}, \sqrt{k})$, where $k = \frac{1}{\text{hidden\_size}}$. As LSTM still suffers from the gradient exploding problem, we set gradient clipping threshold to 10 in our experiments.
- Dropout: We used Dropout after embedding layer and set dropout ratio to 0.5.
- Optimization: We used the AdamW optimizer (Loshchilov and Hutter, 2018) with an initial lr = 0.001 and wd=0.01. The batch size is set to 64.
- Training: We trained model for 20 epochs and choose the best model according to the performance of validation set.

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

Note that, the hyperparameters are consistent across all datasets.

5.2 Evaluation Metrics

We used the micro-F1 score as our main evaluation metrics. micro-F1 (Mi-F1) can be interpreted as a weighted average of the precision and recall. It is calculated globally by counting the total true positives, false positives, and false negatives.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$micro\text{-}F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

5.3 Baselines

- **Binary Relevance (BR)** (Boutell et al., 2004) transforms the task of multi-label classification into the task of binary classification, which is simple and reversible but ignores potential correlations among labels and may lead to the issue of sample imbalance.
- **Label powerset (LP)** (Tsoumakas and Vlahavas, 2007) generates a new class for each possible combination of labels and then solves the problem as a single-label multi-class one.
- **Classifier Chains (CC)** (Read et al., 2011) treats this task as a sequence labeling problem and overcomes the label independence assumption due to classifiers are built upon the previous predictions.
- **CNN** (Kim, 2014) uses multiple convolution kernels to extract text feature, which is then input to the linear transformation layer followed by a sigmoid function to output the probability distribution over the label space.
- **CNN-RNN** (Chen et al., 2017) incorporated CNN and RNN so as to capture both global and local semantic information and model high-order label correlations.
- **SGM** (Yang et al., 2018) **(state-of-the-art)** views the multi-label classification task as a

sequence generation problem, and apply a sequence generation model with a novel decoder structure to solve it.

- **RNN+att** is our implementation of the RNN-based model with the normal attention mechanism.

6 Results

The results of AltXML and baseline models on our IMCM dataset are presented in Table 2. From the results of the conventional baselines, it can be found that the machine-learning-based methods for multi-label text classification still own competitiveness compared with the deep-learning-based methods.

For the generating model, the SGM+GE achieve significant improvements on the IMCM dataset, compared with the machine-learning-based models. However, there is still a certain gap compared with the classification model. By contrast, our proposed model can capture more key features at the same time and achieve the best performance in the evaluation of micro-F1 score, which improves 6.1% of micro-F1 score compared with the SGM+GE.

Model	Mi-P	Mi-R	Mi-F1
BR	76.8	36.8	49.8
CC	70.5	39.9	51.0
LP	50.7	44.9	47.6
SGM+GE	60.6	54.3	57.3
RNN+Att	69.2	57.2	62.6
AltXML	70.0	57.8	**63.3**

Table 2: Results on IMCM Dataset.

We also implement our experiments on public datasets. On the AAPD dataset, similar to the models' performance on the IMCM dataset, our AltXML model achieved good performance, with a 0.8% increase in micro-F1 scores compared to the best, as shown in Table 3.

On the Rcv1v2 dataset, our AltXML model still achieves similar performance on micro-F1 on this dataset comparded with Seq2Seq model (SGM+GE), which illustrates the robustness of our model. Because we have not adjusted the hyperparameters, there is still a lot of space for improvement. The results can be seen in Table 4.

Model	Mi-P	Mi-R	Mi-F1
BR	64.4	64.8	64.6
CC	65.7	65.1	65.4
LP	66.2	60.8	63.4
SGM+GE	74.6	67.5	71.0
RNN+Att	72.0	69.7	70.8
AltXML	71.8	71.9	**71.8**

Table 3: Results on AAPD Dataset.

Model	Mi-P	Mi-R	Mi-F1
BR	90.4	81.6	85.8
CC	88.7	82.8	85.7
LP	89.6	82.4	85.8
CNN	92.2	79.8	85.5
CNN-RNN	88.9	82.5	85.6
SGM+GE	89.7	86.0	**87.8**
RNN+Att	89.1	85.2	87.1
AltXML	90.1	84.6	87.2

Table 4: Results on Rcv1v2 Dataset.

An interesting finding is that, by comparing on three datasets, although the Seq2Seq models achieves the state-of-the-art performance on the Rcv1v2 English dataset, the generalization on our IMCM dataset is insufficient. We think there are two reasons: (1) Compared to the other two datasets, the number of labels for each instance in our dataset is more and there are no obvious semantic boundaries among some labels. (2) Due to the attention mechanism cannot improve the performance of the Seq2Seq model in this task (Lin et al., 2018), Seq2Seq model cannot capture some useful information.

By comparing on the three datasets, our model achieves promising performance.

7 Conclusions

In this paper, we introduce the first Chinese multi-label text classification dataset, IMCM. This dataset focuses on imbalanced multi-label classification. Among many datasets, our model could also give significant improvements over various state-of-the-art baselines. Furthermore, we propose an alternat-

ing attention model to handle the imbalanced problems, and further analysis of experimental results demonstrates that our proposed model not only capture the correlations between labels, but also capture the more features when predicting different labels.

Acknowledgements

This work was supported by Beijing Natural Science Foundation(4192057). We thank anonymous reviewers for their helpful feedback and suggestions.

References

Kush Bhatia, Himanshu Jain, Purushottam Kar, Manik Varma, and Prateek Jain. 2015. Sparse local embeddings for extreme multi-label classification. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 730–738. Curran Associates, Inc.

Matthew R. Boutell, Jiebo Luo, Xipeng Shen, and Christopher M. Brown. 2004. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757 – 1771.

G. Chen, D. Ye, Z. Xing, J. Chen, and E. Cambria. 2017. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2377–2383, May.

Eva Gibaja and Sebastián Ventura. 2015. A tutorial on multilabel learning. *ACM Comput. Surv.*, 47(3):52:1–52:38, April.

Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Himanshu Jain, Yashoteja Prabhu, and Manik Varma. 2016. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *KDD*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *empirical methods in natural language processing*, pages 1746–1751.

David Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397.

Junyang Lin, Xu Sun, Pengcheng Yang, Shuming Ma, and Qi Su. 2018. Semantic-unit-based dilated convolution for multi-label text classification. *empirical methods in natural language processing*, pages 4554–4564.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 115–124. ACM.

Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam. *arXiv: Learning*.

Jinseok Nam, Eneldo Loza Mencía, Hyunwoo J Kim, and Johannes Fürnkranz. 2017. Maximizing subset accuracy with recurrent neural networks in multi-label classification. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5413–5423. Curran Associates, Inc.

Yashoteja Prabhu, Anil Kag, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. 2018. Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 993–1002, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333, Jun.

Yukihiro Tagami. 2017. Annexml: Approximate nearest neighbor search for extreme multi-label classification. pages 455–464. the 23rd ACM SIGKDD International Conference, 08.

Grigorios Tsoumakas and Ioannis Vlahavas. 2007. Random k-labelsets: An ensemble method for multilabel classification. In Joost N. Kok, Jacek Koronacki, Raomon Lopez de Mantaras, Stan Matwin, Dunja Mladenič, and Andrzej Skowron, editors, *Machine Learning: ECML 2007*, pages 406–417, Berlin, Heidelberg. Springer Berlin Heidelberg.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. Sgm: Sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3915–3926. Association for Computational Linguistics.

Ian En-Hsu Yen, Xiangru Huang, Pradeep Ravikumar, Kai Zhong, and Inderjit Dhillon. 2016. Pd-sparse : A primal and dual sparse approach to extreme multiclass and multilabel classification. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *ICML*, volume 48 of *Proceedings of Machine Learning Research*, pages 3069–3077, New York, New York, USA, 20–22 Jun. PMLR.

Fisher Yu and Vladlen Koltun. 2016. Multi-scale context aggregation by dilated convolutions. In *ICLR*.