Hierarchical Label Generation for Text Classification

Jingun Kwon[1,3], **Hidetaka Kamigaito**[1,2], **Young-In Song**[3], **and Manabu Okumura**[1]
[1]Tokyo Institute of Technology
[2]Nara Institute of Science and Technology (NAIST)
[3]Naver Corporation
jingun.kwon@navercorp.com
kamigaito.h@is.naist.jp
song.youngin@navercorp.com
oku@pi.titech.ac.jp

Abstract

Hierarchical text classification (HTC) aims to assign the most relevant labels with the hierarchical structure to an input text. However, handling unseen labels with considering a label hierarchy is still an open problem for real-world applications because traditional HTC models employ a pre-defined label set. To deal with this problem, we propose a generation-based classifier that leverages a Seq2Seq framework to capture a label hierarchy and unseen labels explicitly. Because of no available social media datasets that target at HTC, we constructed a new (**Blog**) dataset using pairs of social media posts and their hierarchical topic labels. Experimental results on the **Blog** dataset showed the effectiveness of our generation-based classifier over state-of-the-art baseline models. Human evaluation results showed that the quality of generated unseen labels outperforms even the gold labels.

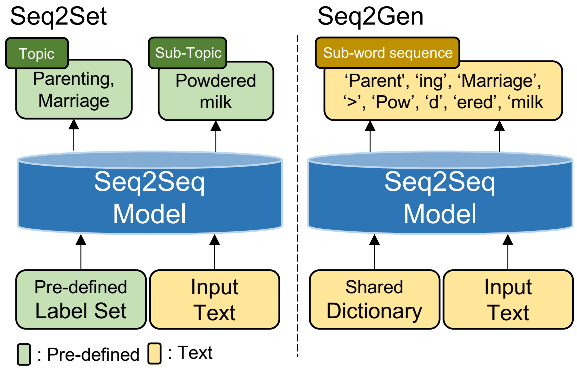

Figure 1: Different from previous Seq2Set (Rivas Rojas et al., 2020), our Seq2Gen can handle unseen labels with sub-word level generation.

1 Introduction

Hierarchical text classification (HTC) aims to assign the most relevant labels with their structure for a given document. Because real-world applications categorize documents into a structured class hierarchy sequence (Silla and Freitas, 2011), such as patent collections (Tikk et al., 2005), web content collections (Dumais and Chen, 2000), and medical record coding (Cao et al., 2020), it is needed to capture the label hierarchy for better categorization.

To solve the HTC task, recent work has focused on enhancing label embeddings with a taxonomic hierarchy (Cao et al., 2020; Zhou et al., 2020; Wang et al., 2021) or considering a sequential classification approach (Rivas Rojas et al., 2020; Yang et al., 2018, 2019) that leverages a Seq2Seq framework to capture the label hierarchy. Despite the previous methods being successful, their approaches classify labels sequentially by choosing them from the pre-defined label set in the training dataset. It is still an open problem for real-world applications to handle unseen labels that do not appear in the pre-defined label set from the training dataset (Banerjee et al., 2019; Aly et al., 2019; Xu et al., 2021). Due to severe deficiencies in annotating data for labels in a hierarchy and handling unseen labels for real-world applications (Liu et al., 2021), we need a general modeling framework for handling unseen labels while explicitly incorporating a label hierarchy to overcome the restriction of the pre-defined label set for the development of real-world text classification applications.

For this purpose, we propose a generation-based classifier that can generate unseen labels in sub-word level. Our method can directly predict labels within a hierarchical structure by considering the label hierarchy as the order of the labels in a sequence. Because all labels are represented as sub-word strings in a shared vocabulary between labels and words, our method can predict unseen labels through generation (Sennrich et al., 2016). To expand unseen labels considerably, we also propose a method to extract knowledge of hierarchical labels from a pre-trained encoder-decoder by semi-supervised learning.

Since there are no available social media datasets

Findings of the Association for Computational Linguistics: EACL 2023, pages 625–632

for HTC, we constructed a new blog dataset in Korean that includes a hierarchical label structure. The dataset contains up to three levels with a document. To evaluate the treatment of unseen labels in detail, we additionally constructed cross-lingual datasets, consisting of Japanese and English social media posts from the Kyoto (Hashimoto et al., 2011) and Reddit (Kim et al., 2019) datasets.

Comparisons between our generation-based and traditional classifiers on the Blog dataset showed that our method outperforms state-of-the-art models for both rank-based and ROUGE metrics. Human evaluation results showed that the quality of our generated unseen labels outperforms even the gold labels. In addition, we confirmed our generation-based classifier can handle unseen labels even on the cross-lingual datasets in a zero-shot setting, that shows the potential for tagging labels with considering a label hierarchy in unseen languages.

2 Problem Formulation

We introduce the task of traditional HTC and formulate how we solve it in our generation-based framework. The traditional HTC has been formalized as choosing labels one-by-one from a predefined label set in the training dataset, for example, with a sequential classification method (Seq2Set). However, handling unseen labels with considering a label hierarchy is important in designing models for real-world applications.

To solve this problem, we formulate the task as topic generation using a Seq2Seq model (Seq2Gen), such as pre-trained BART (Lewis et al., 2020). Figure 1 shows the Seq2Seq framework to generate target labels. It generates labels for an input text as a sequence of label tokens, and thus the label hierarchy can be directly considered through the Seq2Seq model. Because all the labels are represented as sub-word strings in a shared vocabulary between labels and words (Xiong et al., 2021), our model is permitted to generate even unseen labels, that are not included in the predefined label set (Sennrich et al., 2016). Due to the lack of diverse labels with considering their hierarchy in HTC datasets (Kowsari et al., 2017; Sinha et al., 2018), we utilize semi-supervised learning to draw the pre-trained knowledge in the pre-trained Seq2Seq model.

Topic Label	Hierarchical Template
$L = \{l_1\}$	l_1 is a topic.
$L = \{l_1, l_2\}$	l_2 is a sub-topic of l_1.
$L = \{l_1, l_2, l_3\}$	l_2 is a sub-topic of l_1 and l_3 is a sub-topic of l_2.

Table 1: Hierarchical template to map labels into a target topic sequence.

3 Generation-based Classifier

Considering HTC as a language generation task, we use a multi-lingual BART (mBART) (Liu et al., 2020), which is an extended version of a transformer-based pre-trained BART for multiple languages, as our Seq2Seq framework.

3.1 Seq2Seq-based Model

Our generation-based classifier can directly consider a label hierarchy. For learning, we append “>” as a special symbol representing a hierarchy between topics, $\mathbf{L} = \{l_1, l_2, l_3\}$, and concatenate them as a target topic sequence. Let w_i be the i-th token in a document $\mathbf{D} = \{w_1, w_2, ..., w_n\}$. $\mathbf{D}$ is fed into the encoder of the mBART, and then the generated hidden representations with the previous output token, c_{i-1}, are fed into the i-th step of the decoder. Finally, we use the cross-entropy loss between the decoder's output and the label sequence to fine-tune the model, as follows:

$$H^{Enc} = \text{Encoder}(D), \tag{1}$$

$$H^{Dec} = \text{Decoder}(H^{Enc}, c_{i-1}), \tag{2}$$

$$Loss = -\sum_{i \in m} \log(\mathrm{Softmax}(\mathrm{H^{Dec}W + b})), \tag{3}$$

where W and b indicate a learnable weight and bias, respectively, and m indicates the target length.

To show the effectiveness of directly considering a label hierarchy, we additionally consider a template-based Seq2Seq model. For learning, we manually create a hierarchical template, which has slots to map topic labels into a target topic sequence, instead of $\mathbf{L}$. Table 1 shows the hierarchical template to map topics into slots.

3.2 Augmentation with Semi-supervision

Since BART is a pre-trained Seq2Seq model learned with massive text corpora, we assume that we can draw pre-trained knowledge (Petroni et al., 2019) from BART to enhance the label hierarchy and expand labels considerably for dealing with unseen labels. For this purpose, we augment the

Training	Valid	Test
13,705 (1,011)	761 (254)	761 (292)

Table 2: Statistics of Blog. The number in parentheses indicates the number of different labels in each data.

dataset with a *silver* dataset, an automatically annotated dataset by using a model's generation in a manner of semi-supervised learning. As demonstrated by He et al. (2020), we first train a model only with the *silver* dataset, generated by a model trained with the gold dataset, and then fine-tune it with the gold dataset.

4 Blog Dataset

We created a new HTC dataset (Blog) by collecting posts and their topic label sequences from Naver blogs,[1] that contain a large number of different labels compared to the previous HTC datasets (Kowsari et al., 2017; Sinha et al., 2018). The topic label sequences contain up to three hierarchical topic levels. Extracted topic label sequences can be noisy because a blogger can choose only the topic (the top-level class) from 32 classes, and the remaining topic sequence was automatically generated by the Naver blog system. Therefore, we hired experts on social media to annotate a relevance score from 0 to 3 (3 is the best) for a post and its topic label sequence. We filtered posts with scores less than 2 to ensure high quality. Then, we divided them into three parts (training: 90%, valid: 5%, and test: 5%). Table 2 shows the statistics of the created dataset.

To evaluate unseen label generation in cross-lingual few- and zero-shot settings, we additionally created Japanese (Kyoto) and English (Reddit) datasets from publicly available social media post datasets (Hashimoto et al., 2011; Kim et al., 2019). For Kyoto and Reddit, we extracted 249 and 500 posts, respectively. For each post, five human experts annotated a topic label sequence. After preprocessing, we obtained 234 and 400 posts with their label sequences for Kyoto and Reddit, respectively, and divided them into three parts (training: 10%, valid: 5%, and test: 85%). Blog, Kyoto, and Reddit are available upon request.[2]

[1]https://section.blog.naver.com/

[2]Detailed explanations for the datasets are in Appendix A.

5 Experiments

5.1 Experimental Settings

Datasets: Blog, Kyoto, and Reddit were used to compare our generation-based and previous classification methods. To obtain silver data for semi-supervised learning, we additionally extracted 21,520 Naver blog posts. We also evaluated our models on the public HTC dataset, Web of Science (WOS) (Kowsari et al., 2017). It contains 46,985 instances with two levels, where each level consists of 7 and 134 different labels. We divided them into three parts (training: 60%, valid: 20%, and test: 20%).

Evaluation Metrics: Previous studies used a short ranked list of potentially relevant labels to evaluate the classification quality: the precision at top k (P@k) and the Normalized Discounted Cumulative Gain at top k (NDCG@k), where k = 1, 2, 3 (Xun et al., 2020; Zhang et al., 2021). However, these rank-based evaluation metrics could not evaluate the quality of a hierarchical label sequence, and thus, we also used ROUGE-1-F and ROUGE-2-F, that can evaluate the quality of hierarchical label sequences by taking into account label n-grams.

Compared Methods: Our methods are as follows: **Template** uses the proposed hierarchical templates to generate a topic label sequence with mBART.[3] **Seq2Gen** directly generates a topic label sequence with mBART. **Self-Template** and **Self-Seq2Gen** use **Template** and **Seq2Gen** by expanding unseen labels with semi-supervised learning, respectively.

The baselines, which include state-of-the-art models that employ a tree structure of labels, are as follows: **CorNet** utilizes BERT (Devlin et al., 2019) by incorporating a feed-forward layer to consider a label hierarchy (Xun et al., 2020). **MATCH** utilizes BERT by incorporating hypernymy regularization in a loss function to consider hierarchical structures (Zhang et al., 2021).[4] **Seq2Set** is a variant of the state-of-the-art HTC model that sequentially classifies a topic label sequence from a pre-defined label set with mBART. We replaced Bi-GRU with mBART for a fair comparison to our Seq2Gen (Rivas Rojas et al., 2020).

[3]Results using different templates are in Appendix B.

[4]For both CorNet and MATCH, we used a multilingual BERT instead of the original BERT for the cross-lingual setting.

[5]The paired-bootstrap-resampling (Koehn, 2004) was used ($p < 0.05$).

Model	P&N@1	P@2	P@3	N@2	N@3	R1-F	R2-F	Unseen
CorNet	77.79	50.72	36.88	70.03	72.76	47.77	8.76	-
MATCH	78.06	50.72	36.05	70.23	72.10	46.76	9.20	-
Seq2Set	92.38	64.72	43.58	88.36	88.23	81.61	35.50	-
Template	92.12	68.13†	46.25†	89.37	89.44	84.60†	43.17†	91
Seq2Gen	92.25	69.58†	47.39†	89.33	89.53	85.51†	45.42†	102
Self-Template	92.38	68.33	46.30	89.79	89.84	85.88	43.36	74
Self-Seq2Gen	**92.77**	**69.84**	**47.48**	**90.23**	**90.36**	**87.69**‡	**45.95**	62

Table 3: Experimental results on Blog. Unseen indicates the number of different generated unseen labels on the test data. † and ‡ indicate the improvement is significant over the underlined score, respectively.[5]

Model	Kyoto		Reddit	
	R1-F	**R2-F**	**R1-F**	**R2-F**
Few-shot				
CorNet	47.48	15.83	20.60	0.39
MATCH	48.88	18.08	19.64	0.20
Seq2Set	**63.75**	**51.83**	19.69	3.24
Seq2Gen	56.73	35.50	**33.20**	**7.84**
Zero-shot				
Seq2Gen	41.12	13.83	17.48	4.61

Table 4: Results on Kyoto and Reddit.

Model	P&N@1	P@2	N@2	R1-F	R2-F	Unseen
CorNet	78.76	53.89	59.52	53.89	16.78	-
MATCH	74.14	51.07	56.29	51.07	13.53	-
Seq2Set	91.23	85.94	87.14	85.94	80.55	-
Seq2Gen	**91.43**	**86.32**†	**87.48**	**86.32**†	**81.11**†	1

Table 5: Experimental results on WOS. The notations are the same as in Table 3.

5.2 Automatic Evaluation

Table 3 shows the results on Blog. Generating topic labels using the mBART-based models consistently outperformed classifying them using the mBERT-based models. Specifically, the gain was large in the ROUGE metrics. In addition, our generation-based methods, Template and Seq2Gen, outperformed the sequential classifier Set2Set. The proposed Seq2Gen outperformed Template, where the improvement in R2-F was larger than that in R1-F, that indicates Seq2Gen can capture a hierarchical sequence directly compared with the hierarchical template. Moreover, Self-Template and Self-Seq2Gen, that use the *silver* dataset to fine-tune the models, consistently improved the performances. This is because we succeeded in enhancing the label hierarchy with diverse unseen labels. For 21,520 posts in the *silver* dataset, our Seq2Gen could generate 4,385 different unseen labels.

Model	Relevance	Taxonomy	Best
Seq2Set	2.29	2.17	0
Self-Seq2Gen	**2.59**	**2.56**†	23
Gold	2.51	2.46	13

Table 6: Human evaluation results. The notations are the same as in Table 3.

Input: Yoon Restaurant's Kimchi pancake. How to make kimchi pancake, recipe for kimchi pancake. It's been a few days since spring rain has been so moist, so the air is very fresh:) ... **Gold:** Cooking, Recipe **Self-Seq2Gen:** Cooking, Recipe > Kimchi pancake
Input: I can't go to the gym, I can't exercise outside, watch diet YouTube at home. ... The problem with Home Training is that all the exercise moves go by so quickly. ... **Gold:** Health, Medicine **Self-Seq2Gen:** Sports > Home Training

Table 7: Examples of generated unseen labels from Self-Seq2Gen in the Blog dataset.

Table 4 shows the cross-lingual results. The R2-F scores for Seq2Gen, trained with Blog, in the zero-shot setting show that it can generate even cross-lingual unseen labels.[6] Table 5 shows the results on WOS. We can confirm that the generation-based method outperformed the sequential classification method. Thus, our Seq2Gen can work better even for a smaller number of different labels. However, we think the improvements and the number of generated unseen labels are smaller than the ones on Blog due to the smaller number of different labels.

5.3 Human Evaluation and Analysis

We conducted a human evaluation for 50 randomly sampled posts that contain generated unseen labels from our Self-Seq2Gen. Five human annotators graded them with scores from 1 to 3 (3 is the best)

[6]Results including rank-based metrics are in Appendix C.

in terms of Relevance and Taxonomy.[7] We additionally asked the annotators to select the best label sequence from Seq2Set, Self-Seq2Gen, and Gold label sequences. Best indicates the number of cases where the majority among the annotators judged the best. Table 6 shows the human evaluation results. The generated unseen labels from Self-Seq2Gen achieved a higher preference than the Gold labels.

Table 7 shows example generated unseen labels from Self-Seq2Gen. As we expected, our Self-Seq2Gen frequently generated unseen labels with considering the label hierarchy. In the first example, the generated unseen label, "Kimchi pancake", can be considered as a sub-topic of "Cooking, Recipe" because the "Kimchi pancake" is a food name. In the second example, "Home training" can be considered as a sub-topic of "Sports".

6 Conclusion

We proposed a generation-based classifier for HTC. It could handle unseen labels with considering their label hierarchy. In addition, we constructed cross-lingual HTC datasets from social media posts. Automatic evaluation results showed that our generation-based classifier could outperform state-of-the-art models. We confirmed our classifier could handle unseen labels by human evaluation.

7 Ethical Considerations

We created the new datasets of `Blog`, `Kyoto`, and `Reddit` for the HTC task. The created datasets have been collected in a manner which is consistent with the terms of use of any sources and the intellectual property and privacy rights of the original authors of the texts. Please note that we have confirmed by our legal team and the datasets will be available upon request for only research purpose.

8 Limitations

Although our Seq2Gen could generate unseen labels on cross-lingual datasets in the zero-shot setting, that shows the potential of tagging labels with considering their label hierarchy, it was difficult to outperform the few-shot setting. In the future, we plan to incorporate cross-lingual label trees for the zero-shot setting.

[7] Relevance and Taxonomy indicate how much the generated label sequences are related to the input context and the quality of the label hierarchy, respectively.

References

Rami Aly, Steffen Remus, and Chris Biemann. 2019. Hierarchical multi-label classification of text with capsule networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 323–330, Florence, Italy. Association for Computational Linguistics.

Siddhartha Banerjee, Cem Akkaya, Francisco Perez-Sorrosal, and Kostas Tsioutsiouliklis. 2019. Hierarchical transfer learning for multi-label text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6295–6300, Florence, Italy. Association for Computational Linguistics.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020. HyperCore: Hyperbolic and co-graph representation for automatic ICD coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3105–3114, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Susan Dumais and Hao Chen. 2000. Hierarchical classification of web content. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, page 256–263, New York, NY, USA. Association for Computing Machinery.

Chikara Hashimoto, Sadao Kurohashi, Daisuke Kawahara, Keiji Shinzato, and Masaaki Nagata. 2011. Construction of a blog corpus with syntactic, anaphoric, and sentiment annotations. *Journal of Natural Language Processing*, 18(2):175–201.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2020. Revisiting self-training for neural sequence generation. In *International Conference on Learning Representations*.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive summarization of Reddit posts with multi-level memory networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2519–2531, Minneapolis, Minnesota. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural*

Language Processing. Association for Computational Linguistics.

Kamran Kowsari, Donald E. Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S. Gerber, and Laura E. Barnes. 2017. Hdltex: Hierarchical deep learning for text classification. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 364–371.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Hui Liu, Danqing Zhang, Bing Yin, and Xiaodan Zhu. 2021. Improving pretrained models for zero-shot multi-label text classification through reinforced label hierarchy reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1051–1062, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Kervy Rivas Rojas, Gina Bustamante, Arturo Oncevay, and Marco Antonio Sobrevilla Cabezudo. 2020. Efficient strategies for hierarchical text classification: External knowledge and auxiliary tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2252–2257, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Carlos N Silla and Alex A Freitas. 2011. A survey of hierarchical classification across different application domains. In *Data Mining and Knowledge Discovery*, 22(1-2):31–72, New York, NY, USA.

Koustuv Sinha, Yue Dong, Jackie Chi Kit Cheung, and Derek Ruths. 2018. A hierarchical neural attention-based text classifier. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 817–823, Brussels, Belgium. Association for Computational Linguistics.

Domonkos Tikk, György Biró, and Jae Dong Yang. 2005. *Experiment with a Hierarchical Text Categorization Method on WIPO Patent Collections*, pages 283–302. Springer US, Boston, MA.

Xuepeng Wang, Li Zhao, Bing Liu, Tao Chen, Feng Zhang, and Di Wang. 2021. Concept-based label embedding via dynamic routing for hierarchical text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5010–5019, Online. Association for Computational Linguistics.

Yijin Xiong, Yukun Feng, Hao Wu, Hidetaka Kamigaito, and Manabu Okumura. 2021. Fusing label embedding into BERT: An efficient improvement for text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1743–1750, Online. Association for Computational Linguistics.

Linli Xu, Sijie Teng, Ruoyu Zhao, Junliang Guo, Chi Xiao, Deqiang Jiang, and Bo Ren. 2021. Hierarchical multi-label text classification with horizontal and vertical category correlations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2459–2468, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Guangxu Xun, Kishlay Jha, Jianhui Sun, and Aidong Zhang. 2020. Correlation networks for extreme multi-label text classification. KDD '20, page 1074–1082, New York, NY, USA. Association for Computing Machinery.

Pengcheng Yang, Fuli Luo, Shuming Ma, Junyang Lin, and Xu Sun. 2019. A deep reinforced sequence-to-set model for multi-label classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5252–5258, Florence, Italy. Association for Computational Linguistics.

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yu Zhang, Zhihong Shen, Yuxiao Dong, Kuansan Wang, and Jiawei Han. 2021. Match: Metadata-aware text classification in a large hierarchy. In *WWW'21*, pages 3246–3257. ACM / IW3C2.

Jie Zhou, Chunping Ma, Dingkun Long, Guangwei Xu, Ning Ding, Haoyu Zhang, Pengjun Xie, and Gongshen Liu. 2020. Hierarchy-aware global model for hierarchical text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1106–1117, Online. Association for Computational Linguistics.

A 32 Topics for Naver blog System

Table 8 shows 32 topic classes (top-level) from Naver blog system.

For Kyoto and Reddit, to establish the same setting as for Blog, the experts first annotated the topic label (the top-level class) from given 32 classes. Then, they annotated hierarchical label sequences up to three-levels if they consider subsequent labels are required. We deleted posts with no majority for the topic label. We obtained 234 and 400 posts with their label sequences for Kyoto and Reddit, respectively, and divided them into three parts (training: 10%, valid: 5%, and test: 85%).

For Reddit and Kyoto, each input text is not one-to-one matching for target labels, which is different from the **Blog** dataset. For training, we considered all different target label sequences. For the evaluation, we selected maximized scores by regrading them as multiple references. To assess the agreement between the participants for the datasets, we used Fleiss' Kappa (L. Fleiss, 1971). We obtained Kappa scores of 0.55 for Kyoto and 0.23 for Reddit, indicating moderate and fair agreements, respectively.

B Results using different templates.

We study the various manually created hierarchical templates using valid Blog because different hierarchical templates can express the same meaning. Table 9 shows the performance using different templates. On the basis of the valid results in terms of average ROUGE-F scores, we use the top performing template in our experiments.

C Results on Kyoto and Reddit datasets

Table 10 includes both rank-based and ROUGE metrics on Kyoto and Reddit.

	Topic
1	Literature, Book
2	Movie
3	Art, Design
4	Performance, Exhibition
5	Music
6	Drama
7	Star, Celebrity
8	Cartoon, Anime
9	Broadcast
10	Everyday, Thoughts
11	Parenting, Marriage
12	Pet, Companion animal
13	Good article, Image
14	Fashion, Beauty
15	Interior, DIY
16	Cooking, Recipe
17	Product review
18	Horticulture, Cultivation
19	Game
20	Sports
21	Picture
22	Car
23	Hobby
24	Domestic travel
25	World travel
26	Restaurant
27	IT, Computer
28	Society, Politics
29	Health, Medicine
30	Business, Economy
31	Language, Foreign language
32	Education, Academic

Table 8: 32 topics from Blog datasets.

Topic Label	Hierarchical Template	R1-F	R2-F	Avg R-F
$L = \{l_1\}$ $L = \{l_1, l_2\}$ $L = \{l_1, l_2, l_3\}$	l_1 is a topic. l_2 is a sub-topic of l_1. l_2 is a sub-topic of l_1 and l_3 is a sub-topic of l_2.	86.92	**45.90**	**66.41**
$L = \{l_1\}$ $L = \{l_1, l_2\}$ $L = \{l_1, l_2, l_3\}$	l_1 is a topic. l_1 is a topic and l_2 is a sub-topic of l_1. l_1 is a topic, l_2 is a sub-topic of l_1, and l_3 is a sub-topic of l_2.	86.72	44.88	65.80
$L = \{l_1\}$ $L = \{l_1, l_2\}$ $L = \{l_1, l_2, l_3\}$	l_1 is a topic. l_1 is a parent topic of l_2. l_1 is a parent topic of l_2 and l_2 is a parent topic of l_3.	**87.31**	43.82	65.57
$L = \{l_1\}$ $L = \{l_1, l_2\}$ $L = \{l_1, l_2, l_3\}$	l_1 is a topic. l_1 is a topic and l_1 is a parent topic of l_2. l_1 is a topic, l_1 is a parent topic of l_2, and l_2 is a parent topic of l_3.	85.87	44.20	65.04

Table 9: Results using different hierarchical templates.

Model	**Kyoto**							**Reddit**						
	P&N@1	**P@2**	**P@3**	**N@2**	**N@3**	**R1-F**	**R2-F**	**P&N@1**	**P@2**	**P@3**	**N@2**	**N@3**	**R1-F**	**R2-F**
Few-shot														
CorNet	54.50	56.50	41.67	55.66	54.61	47.48	15.83	35.59	22.79	17.25	27.03	26.59	20.60	0.39
MATCH	57.50	56.75	43.00	56.92	56.65	48.88	18.08	29.71	20.44	15.78	24.02	25.68	19.64	0.20
Seq2Set	64.00	**69.75**	**46.50**	**68.45**	**67.93**	**63.75**	**51.83**	37.06	21.91	14.71	26.82	26.10	19.69	3.40
Seq2Gen	**65.50**	62.50	**46.50**	62.92	60.22	56.43	35.08	**51.18**	**36.76**	**24.71**	**41.70**	**40.01**	**33.20**	**7.84**
Zero-shot														
Seq2Gen	28.00	42.00	28.00	41.73	41.73	41.12	13.83	22.53	15.29	10.20	21.58	21.58	17.48	4.61

Table 10: Evaluation results on the **Kyoto** and **Reddit** datasets.