Semantic-Unit-Based Dilated Convolution for Multi-Label Text Classification

Junyang Lin[1,2] Qi Su[1] Pengcheng Yang[2] Shuming Ma[2] Xu Sun[2]

School of Foreign Languages, Peking University

MOE Key Laboratory of Computational Linguistics, Peking University

{linjunyang, sukia, yang_pc, shumingma, xusun}@pku.edu.cn

Abstract

- **A novel model for multi-label text classification based on Seq2Seq;**
- **Generate semantic unit representations with dilated convolution;**
- **Hybrid attention to integrate semantic unit and word information;**
- **Improved results on benchmark datasets**
- **Comparable to hierarchical models but with fewer costs.**

Motivation

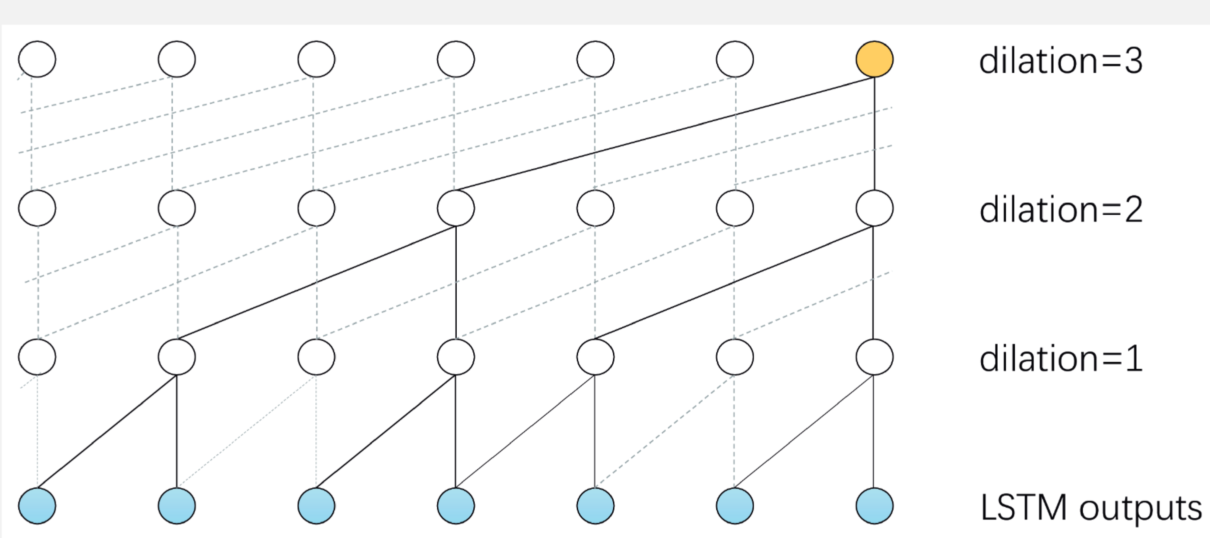
- **Labels for text have internal correlation;**
- **The relation between label and text is complex;**
- **Semantic unit demonstrating event is more informative;**
- **Significant key words can make a difference.**

Sequence-to-Sequence as Baseline

- **Encoder: Bidirectional LSTM**
- **Decoder: LSTM for sequential decoding. Training is with teacher forcing.**
- **Attention mechanism: global attention for the relevant source-side information**

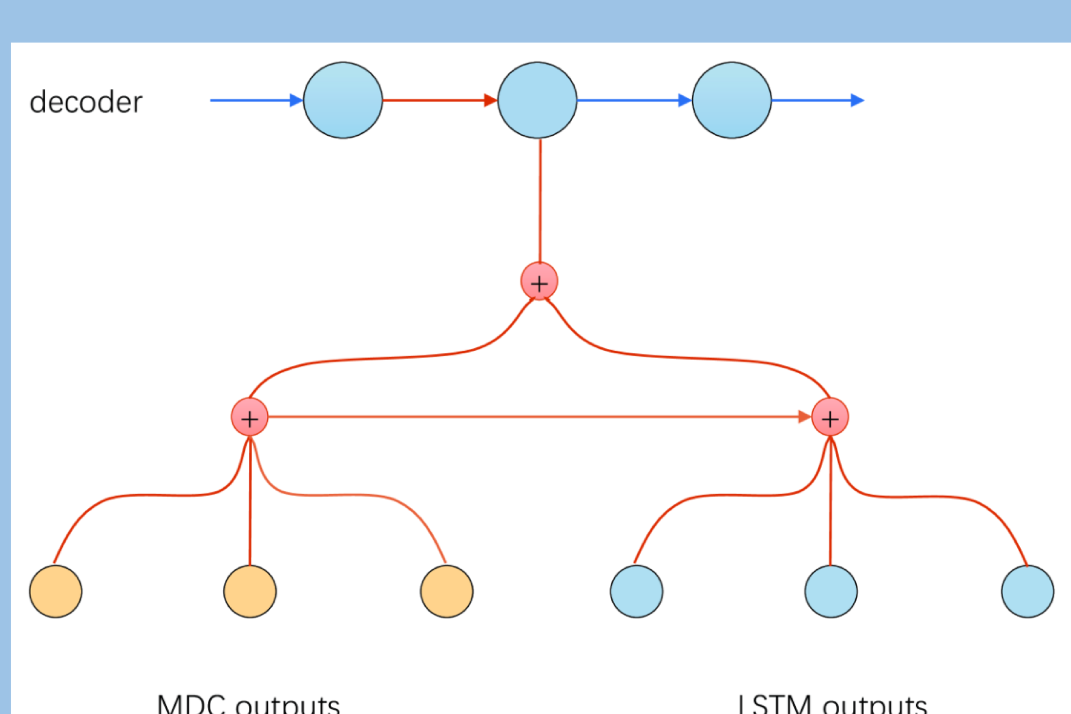
Dilated Convolution for Semantic Unit

- **Multi-level dilated convolution over the outputs of the encoder tend to extract information of semantic units.**



Hybrid Attention for Integration

- **Hybrid Attention allows the integration of the attention to the semantic unit information and the word information.**



Experiments

- **Datasets: Reuters Corpus Volume I (RCV1-v2) and Ren-CECps;**
- **Metric: Hamming Loss and MicroF1 score**

$$HL = \frac{1}{L}\sum \mathbb{I}(y \neq \hat{y})$$

$$microF_1 = \frac{\sum_{j=1}^{L} 2tp_j}{\sum_{j=1}^{L} 2tp_j + fp_j + fn_j}$$

Results

Models	HL(-)	P(+)	R(+)	F1(+)
BR	0.0086	0.904	0.816	0.858
CC	0.0087	0.887	0.828	0.857
LP	0.0087	0.896	0.824	0.858
CNN	0.0089	**0.922**	0.798	0.855
CNN-RNN	0.0085	0.889	0.825	0.856
S2S	0.0082	0.883	0.849	0.866
S2S+Attn	0.0081	0.889	0.848	0.868
Our Model	**0.0072**	0.891	**0.873**	**0.882**

Table 2: Performance on the RCV1-V2 test set. HL, P, R, and F1 denote hamming loss, micro-precision, micro-recall and micro-F_1, respectively ($p < 0.05$).

Models	HL(-)	P(+)	R(+)	F1(+)
BR	**0.1663**	**0.649**	0.472	0.546
CC	0.1828	0.572	0.551	0.561
LP	0.1902	0.556	0.517	0.536
CNN	0.1726	0.628	0.512	0.565
CNN-RNN	0.1876	0.576	0.538	0.556
S2S	0.1814	0.587	0.571	0.579
S2S+Attn	0.1793	0.589	0.573	0.581
Our Model	0.1782	0.593	**0.585**	**0.590**

Table 3: Performance of the models on the Ren-CECps test set. HL, P, R, and F1 denote hamming loss, micro-precision, micro-recall and micro-F_1, respectively ($p < 0.05$).

Ablation tests and Comparison with Hierarchical Models

Models	HL(-)	P(+)	R(+)	F1(+)
w/o attention	0.0086	**0.904**	0.816	0.871
attention	0.0087	0.887	0.828	0.869
MDC	0.0074	0.889	0.871	0.880
additive	0.0073	0.888	0.871	0.879
hybrid	**0.0072**	0.891	**0.873**	**0.882**

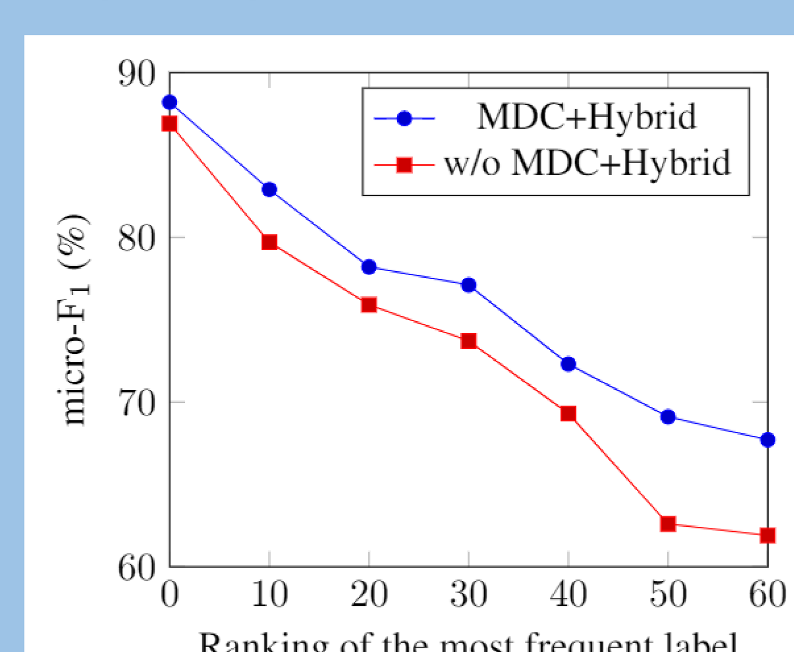
Table 4: Performance of the models with different attention mechanisms on the RCV1-V2 test set. HL, P, R, and F1 denote hamming loss, micro-precision, micro-recall and micro-F_1, respectively ($p < 0.05$).

Models	HL(-)	P(+)	R(+)	F1(+)
Hier-5	0.0075	0.887	0.869	0.878
Hier-10	0.0077	0.883	0.873	0.878
Hier-15	0.0076	0.879	0.879	0.879
Hier-20	0.0076	0.876	**0.881**	0.878
Our model	**0.0072**	**0.891**	0.873	**0.882**

Table 5: Performance of the hierarchical model and our model on the RCV1-V2 test set. Hier refers to hierarchical model, and the subsequent number refers to the length of sentence (word) for sentence-level representations ($p < 0.05$).

Performance on labels of low frequency

- **Remove the top 10, 20, 30, 40, 50 and 60 most frequent labels subsequently**
- **More robust to the classification of labels of low frequency**



Conclusion

- **A new model for multi-label text classification with the combination of Seq2Seq and dilated convolution;**
- **Classification based on semantic units and key words;**
- **Outperform the baselines and robust to labels of low frequency.**