УДК 004.855.5

MULTI-LABEL TEXT CLASSIFICATION FOR LEGAL DOCUMENTS

Шахин Зейн
(Университет ИТМО)
Научный руководитель: к.т.н., доцент Муромцев Дмитрий Ильич
(Университет ИТМО)

Multi-label text classification is a challenging task in natural language processing, this task includes assigning subset of labels to a given document. What is challenging about this task is the big number of classes, e.g. Wikipedia dataset is annotated with hundreds of thousands of tags. Another challenge is that the labels follows power-law distribution. Legal documents often come in the form of long texts; however most current state of the art models deal only with fixed context. In this research, we aim at improving the current state of the art on this task and exploiting hierarchical information to enhance the quality of the model.

Two approaches to deal with Multi-Label Text Classification MLTC tasks, classification based MLTC and ranking-based models.

1. Ranking-based models

classes are ranked based on the probability of assigning them to a given document, then top k classes are selected as labels for this document. There are different approaches to achieve this goal, using BM25 algorithm, a profile for each class is extracted by concatenating all titles assigned with this class [1], after that tf-idf vector representation is calculated based on this profile. During inference tf-idf vector representation is calculated for the input document, and rank of a class is calculated as cosine similarity between the input vector representation and the vector representation related to this class.

2. Classification based methods

machine learning methods: such as support vector machine and binary regression use n-grams and bag of words BOW or tf-idf vector of them as input features, and a model is trained to maximize the likelihood of getting the correct labels given a document's features representation. This feature representation is sparse. These methods doesn't preserve sequential information and it doesn't use them during training.

Deep learning methods: Recurrent Neural Networks are used for sequence modeling, alongside with gated cells such as LSTM and GRU to capture long range dependency. RNNs suffers from vanishing gradients which limits the used context to 200 tokens on average [2,3]. Convolution neural networks use filters to calculate relations between consequence tokens, and usually used alongside with RNN above it.Transformer based models doesn't use any recurrence or convolution in its structure, it includes multi-head attentions and all current state of the art models are based on Transformer [4].

References

1. Khandelwal, U., He, H., Qi, P., & Jurafsky, D. (2018). Sharp nearby, fuzzy far away: How neural language models use context. ArXiv Preprint ArXiv:1805.04623.
2. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V, & Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. ArXiv Preprint ArXiv:1901.02860.
3. Khandelwal, U., He, H., Qi, P., & Jurafsky, D. (2018). Sharp nearby, fuzzy far away: How neural language models use context. ArXiv Preprint ArXiv:1805.04623.
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998–6008).