README

Abstract

This README document contains information about other files in a directory "software_data".

1 The information of the directory "software"

We implement our proposed neural network model SPANSEGTAG based on the public implementation using Python 3 language programming[1] of the research (Tian et al., 2020b). In this section, we describe all files and directories in the directory "software_data". For each code file, we have a comment referring to the public original code or paper.

1.1 The "pytorch_pretrained_bert" directory

This directory contains original code files from the research of Tian et al. (2020b).

1.2 The "pytorch_pretrained_zen" directory

This directory contains original code files from the research of Tian et al. (2020b).

1.3 The "TwASP_prediction" directory

This directory contains all prediction files of best models[2] from the research of Tian et al. (2020a). We can not reproduce results on UD2 dataset. For each file in this directory, its file name has format "{dataset_name} _ {encoder_name} _ test.txt", where dataset_name in {CTB5, CTB6, CTB7, CTB9, UD1} and encoder_name in {BERT, ZEN}.

1.4 The "our_prediction" directory

Since our pre-trained models can not be included in supplementary *.zip archive file, therefore we include only prediction files of our models in the "our_prediction" directory for you reference. This directory contains all prediction files of all models of our SPANSEGTAG. For each file in this directory, its file name has format "{dataset_name} _ {encoder_name} _ {dimension_of_MLPs} _ {dataset_type}.txt", where dataset_name in {CTB5, CTB6, CTB7, CTB9, UD1, UD2} and encoder_name in {BERT}, dimension_of_MLPs in {100, 200, 300, 400, 500} and dataset_type in {dev, test}.

1.5 All files in the "software" directory

- wmseg_main.py: The main file of our source code. You can see the "1_train_test.py" with example running code.
- wmseg_model.py: Contain our neural network model in our research. You can find "span_decode" function in "WMSeg" class, this function refers to the SPANPOSTPROCESSOR in our paper.
- wmseg_helper.py: Contain vocabulary processor.
- wmseg_eval.py: Contain evaluation functions.
- 1_train_test.py: Example training and testing code.
- 2_datasets_statistics.py: To produce Table 1 in our paper.
- 3_significance_test.py: To find significant level in Table 3 of our paper.
- 4_recall_of_out-of-vocabulary_and_in-vocabulary_words.py: Table 4 in our paper.
- 5_combination_ambiguity_string_error.py: Table 5 in our paper.

[1] https://github.com/SVAIGBA/WMSeg

[2] https://github.com/SVAIGBA/TwASP/tree/master/models

2 The information of the directory "data"

The directory "data" contains six other directories of datasets that we used in our research: CTB5, CTB6, CTB7, CTB9, UD1, and UD2. In each dataset directory, it contains three files in *.tsv format of train, dev, and test dataset.

References

Yuanhe Tian, Yan Song, Xiang Ao, Fei Xia, Xiaojun Quan, Tong Zhang, and Yonggang Wang. 2020a. Joint Chinese Word Segmentation and Part-of-speech Tagging via Two-way Attentions of Auto-analyzed Knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8286–8296. Association for Computational Linguistics.

Yuanhe Tian, Yan Song, Fei Xia, Tong Zhang, and Yonggang Wang. 2020b. Improving Chinese Word Segmentation with Wordhood Memory Networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8274–8285, Online. Association for Computational Linguistics.