NLPCC-2020 Shared Task on AutoIE Call for Participant

ZhuiYi Technology

Background

Entity extraction is a fundamental problem in language technology. Most previous work focus on the scenario in which labelled data is provided for interested entities. However, the categories of entities can be hierarchical and cannot be enumerated sometimes. Thus, a generic solution cannot depend on the hypothesis that enough labeled data is provided.

Task

This task is to build IE systems with Noise and Incomplete annotations. Specifically, given a list of entities of specific type and an unlabeled corpus containing these entities, the task aims to build an IE system which may recognize and extract the interested entities of given types. The task setting is very practical and thus the proposed solutions may generalize well in real world applications.
Note:

1. entity is a extended concept of named entity in this task. Some words without a specific name are also very important for downstream applications, therefore, they are included in this information extraction task
2. No human annotation and correction are allowed for train and test dataset.
3. Dev dataset with full label may be used in the training step in any way.

Data

The corpus are from caption text of YouKu video. Three categories of information are considered in this task, which are TV, person and series. All data are split into 3 datasets for training, developing and testing.

Train dataset
1. Unlabelled corpus containing 10000 samples, the entities are labelled by string matching with the given entity lists.
2. Entity lists with specific category, which may cover around 30% of entities appearing in the unlabelled corpus

Dev dataset : 1000 samples with full label
Test dataset : 2000 samples with full label

Submission & Evaluation

For submission, please write the prediction result into a single file and email it to Xuefeng Yang（杨雪峰） email：ryan@wezhuiyi.com

The submission file format should be the same as the format of given dev dataset. To be specific, each sample is separated by a blank line and each char in sample is labelled by BIE format. All labels are B-TV, I-TV, E-TV, B-PER, I-PER, E-PER, B-NUM, I-NUM, E-NUM, and 0.

For evaluation. all the system will be evaluated against 2000 test samples with full annotation. Ranking of submissions are based on the accuracy of these test samples.

A eval.py script is provided to calculate the accuracy and valid prediction format.

Prizes

This task will award prizes for top 3 teams. Winners will get the award certificates issued by NLPCC and CCF Technical Committee on Chinese Information Technology.

First prize: 5000 RMB + award certificate
Second prize: 3000 RMB + award certificate
Third prize: 1000 RMB + award certificate

Website

Further arrangement and baseline system will be published in https://github.com/ZhuiyiTechnology/AutoIE.

Organizers:

Xuefeng Yang (ZhuiYi Technology)
email: ryan@wezhuiyi. com
Benhong Wu (ZhuiYi Technology)
email: wubenhong@wezhuiyi.com

Zhanming Jie (Singapore University of Technology and Design)
email: zhanming_jie@mymail.sutd.edu.sg