NLPCC 2021 Shared Task Guideline: AutoIE2

Background

Sub-events identification is a very fundamental problem in the field of information extraction, especially in emergency situations (e.g., terrorist attacks). It is challenging for two reasons:

1. Data confusing and imbalance. Events usually evolve rapidly and successive sub-events occur. Only a few target sub-events data need to be identified from the large volume of events related data.
2. Low resource. Usually only a limited amount of labelled seed data is given for learning and more annotating datasets are expensive and time consuming.

However, the existing works cannot fully meet the requirements, and thus better few shot learning and data selection models for sub-event identification are crucial.

Task

The goal of this task is to build an IE system (Information Extraction system) that can quickly adapt to a new occurring sub-event. Specifically, there are two settings of this task:

1. Given a large number of event-related corpus and a few labelled seed data, the task aims to build an IE system which can identify the target sub-events.
2. Besides designing machine learning models, annotating data selected from the unlabelled corpus is also allowed, but the size of the labelled data from the unlabelled corpus is fixed. How to select the best data to annotate and enrich training dataset is also an important step in this task.

The task settings are very practical; thus the proposed solutions may generalize well in real world applications.

Note:

1. Three sub-events are included in the task and will be released with the seed dataset.
2. Human annotation and correction are allowed for training dataset which is composed of seed set and data annotated by participants from unlabeled corpus.
3. The size of data annotated by participants may not exceed 100 per sub-event .

Data

All corpus provided are obtained from comments (generally 8 to 120 characters long). The corpus is split into three parts, i.e., unlabeled dataset, seed dataset and testing dataset. The labelled seed dataset (100 samples per event) and unlabeled dataset (100K for 3 events) are released to participants to construct their own training set and developing set, and the testing dataset (around 2k per event) is used for final evaluation.

More details about these three datasets are as follows:

- ✧ Unlabeled dataset: totally 100,000 samples related to the three sub-events.
- ✧ Seed dataset: 100 labeled samples per sub-event.
- ✧ Test dataset : 2000 labeled samples per sub-event.

Note: Test set will be released on 2021/6/5.

Submission & Evaluation

For submission, please write the prediction result into a single file and email it to Xingyu Bai bxy20@mails.tsinghua.edu.cn

There are two settings for this evaluation task and the final evaluation is the average accuracy of two settings.

- S1: few sample problem setting: the size of training data may not exceed 100 per sub-event, human annotation is not allowed.
- S2: data selection problem setting: human annotation is allowed. The size of training data which is composed of seed set and data annotated from unlabeled corpus may not exceed 200.

The format of submission file should be the same as the format of given seed dataset. To be specific, each sample in the test dataset is labelled by 3, 2, 1 and 0.

An eval.py script is provided to calculate the accuracy and verify prediction format.

Prizes

This task will provide 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 arrangements and the baseline system will be published in https://github.com/IIGROUP/AutoIE2

Organizers:

Xuefeng Yang
email: yang0302@e.ntu.edu.sg

Weigang Guo
email: springwg@163.com

Xinyu Bai
email: bxy20@mails.tsinghua.edu.cn

Yujiu Yang
email: yang.yujiu@sz.tsinghua.edu.cn