

SereTOD Track1: Information

Extraction from dialog transcripts

Team 仓颉

Abstract

We mostly employed the following techniques:

1, The domain knowledge enhancement:

MLM and NSP post-pretrain utilizing label and unlabeled data, promote 1%–2%

2, Application of context information:

We experimented with a few different ways to incorporate context into the extraction models. Task 1 did not see any advancements.

The usage of context in the classification job can advance 1–2% for the Separate Entity Type Prediction Model in Task 2.

3, Slot Task Modeling Scheme:

The benefits of several technical schemes are combined in multi-model results fusion voting employing sequence-tagging and span dual technological scheme.

System Pipeline

We reuse the procedure of the pipeline that the officially provided, and we trained a post pretraining model additionally.

1. Train post-pretrain model
2. Entity Extraction
3. Entity Coreference
4. Slot Filling
5. Entity Slot Alignment

1. Train post-pretrain model

Data:

- 1、 10,000 dialogue labeled data official provided.
- 2、 90,000 dialogue unlabeled data build pretrain model train data.

Task:

- 1、 mask language model
- 2、 next sentence prediction

```
@inproceedings{cui-etal-2020-revisiting,  
  title = "Revisiting Pre-Trained Models for {C}hinese Natural Language  
Processing",  
  author = "Cui, Yiming and  
  Che, Wanxiang and  
  Liu, Ting and  
  Qin, Bing and  
  Wang, Shijin and
```

```
    Hu, Guoping",
    booktitle = "Proceedings of the 2020 Conference on Empirical Methods
in Natural Language Processing: Findings",
    month = nov,
    year = "2020",
    address = "Online",
    publisher = "Association for Computational Linguistics",
    url = "https://www.aclweb.org/anthology/2020.findings-emnlp.58",
    pages = "657--668",
}
```

2. Entity Extraction

Data:

- 1、 10,000 dialogue labeled data official provided.
- 2、 Data augmentation based on labeled data.

Task:

- 1、 base model 1: post-pretrain mac-bert + CRF.
 - a、 Use BIOES labeled word sequence.
 - b、 Normalize digital numerical representation to patterns.
- 2、 base model 2: post-pretrain mac-bert + Span.
- 3、 base model 3: the baseline model (Entity Extraction) officially provided (but we use BIOE label system, mac-bert-large pretrain model, and fix prediction results boundary)

4、 base model 4: a single classify model to predict ents–type.

5、 vote results: Ensemble 1、 2、 3、 4、 these four single models for voting

```
@article{xu2020cluener2020,  
  title={CLUENER2020: Fine-grained Name Entity Recognition for Chinese},  
  author={Xu, Liang and Dong, Qianqian and Yu, Cong and Tian, Yin and Liu,  
Weitang and Li, Lu and Zhang, Xuanwei},  
  journal={arXiv preprint arXiv:2001.04351},  
  year={2020}  
}
```

3. Entity Coreference

Data:

1、 10,000 dialogue labeled data.

Task:

1、 reuse baseline model that official provided

4. Slot Filling

Data:

1、 10,000 dialogue labeled data official provided.

2、 Data augmentation based on labeled data.

Task:

Same as 2. Entity Extraction

5. Entity Slot Alignment

Data:

1、 10,000 dialogue labeled data.

Task:

1、 Reuse baseline model that official provided

```
@article{ou2022achallenge,  
title={A Challenge on Semi-Supervised and Reinforced Task-Oriented Dialog  
Systems},  
author={Zhijian Ou and Junlan Feng and Juanzi Li and Yakun Li and Hong  
Liu and Hao Peng and Yi Huang and Jiangjiang Zhao},  
journal={arXiv preprint arXiv:2207.02657},  
year={2022}  
}
```

```
@article{Liu2022InformationEA,  
title={Information Extraction and Human-Robot Dialogue towards Real-life  
Tasks: A Baseline Study with the MobileCS Dataset},  
author={Hong Liu and Hao Peng and Zhijian Ou and Juan-Zi Li and Yi Huang  
and Junlan Feng},  
journal={arXiv preprint arXiv:2209.13464},  
year={2022}  
}
```