Contents

1 Notes from *<2021-02-02 Tue>*

1.1 Present at meeting

, - [X] Peter , - [X] Sudarshan , - [X] Kalpesh , - [X] Ari , - [X] Sweta , - [X] Mike

1.2 Agenda

, - Comments and corrections to last meting notes , - Reports from the sub teams , - Discussion , - Final round

1.3 Notes

, . . . ,

2 Goal: Publish a paper

3 Topic: NLP Data Augmentation

3.1 Approach

- Find existing examples within large corpus such as Common Crawl using embedding
- Generating from scratch(May not efficient)
- Modifying existing data(Style Transfer, . . .)

4 Focus on Classification Tasks

4.1 Ari's suggestion:

- Intent Classification on banking domain (Open a bank account)
- Sentiment Classification(3 classes)

4.2 Kalpesh:

- Citation intent classification
- Topic Classification

5 Evaluation:

- Held-out Test Accuracy
- Adversarial held-out accuracy
- Checklist Evaluation for Linguistic Generalization
- Kalpesh: 10-15 Test Datasets for generalization - Which paper?

6 Dataset: IMDB

7 TODO Create a Jupyter Notebook for Baseline

- Sentiment Classification on IMDB

8 Shared Vector Embedding space is enormous. Focus on a small subset.

- Linguistic phenomena
- TF-IDF
- Prefix String
- Kalpesh: Dense Embedding(DPR)

9 Will discuss further after setting up toy retrieval pipeline

10 Oracle's previous approach

- Collecting the farthest embedding.
- May collect irrelevant data. Thus, think about other way

- e.g. Collect task-relevant examples and take the farthest embedding

11 Kalpesh suggested Don't stop Pretraining embedding an additional pretraining with relevant topic

-

12 Get better semantic representation(embedding) by pre-training with relevant domain corpus

- ROBERTA-DAPT and ROBERTA-TAPT are available online.

`https://huggingface.co/allenai`

13 Ari suggested two papers on data augmentation strategy

14 Additional Papers

14.1 VAMPIRE(`https://github.com/allenai/vampire`)

- This is the augmentation method for the previous paper "don't stop pretraining"
- Got it working without any issue.

15 ACTIONS

, This is the general list of Actions

15.1 TODO Action #4 Talk to companies #4:Sam

15.2 TODO Action #5 foo #5:Me