INTENT INDUCE WITH SCCL AND OUT-DOMAIN DATA AUGMENTATION

Jun Gao

<2022-11-04 >

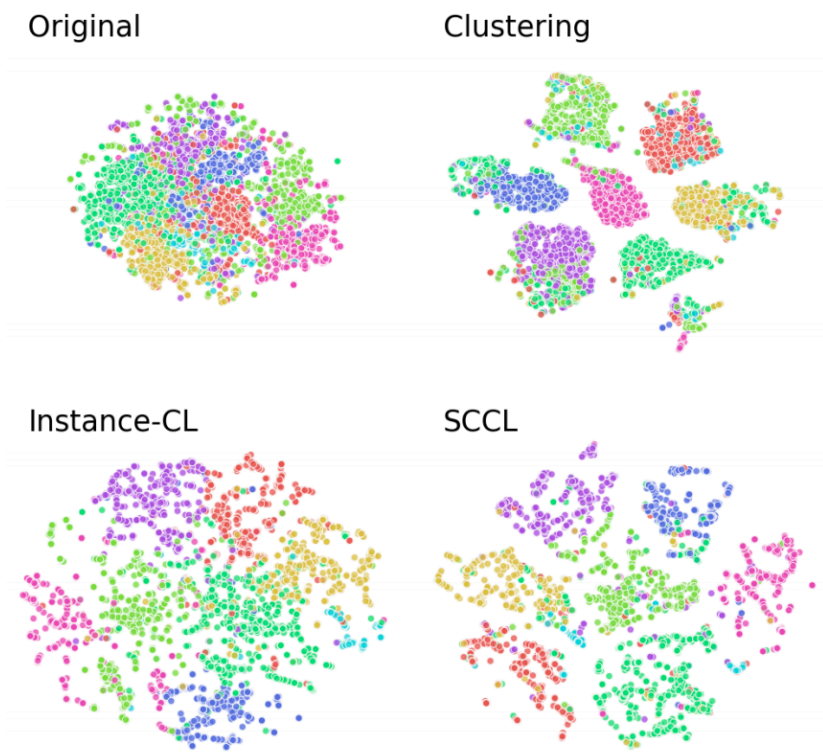
Outline

dstc11

https://drive.google.com/file/d/
1itlby2Ypq3sRVtOY1alr3ygjPZZdB2TT/view

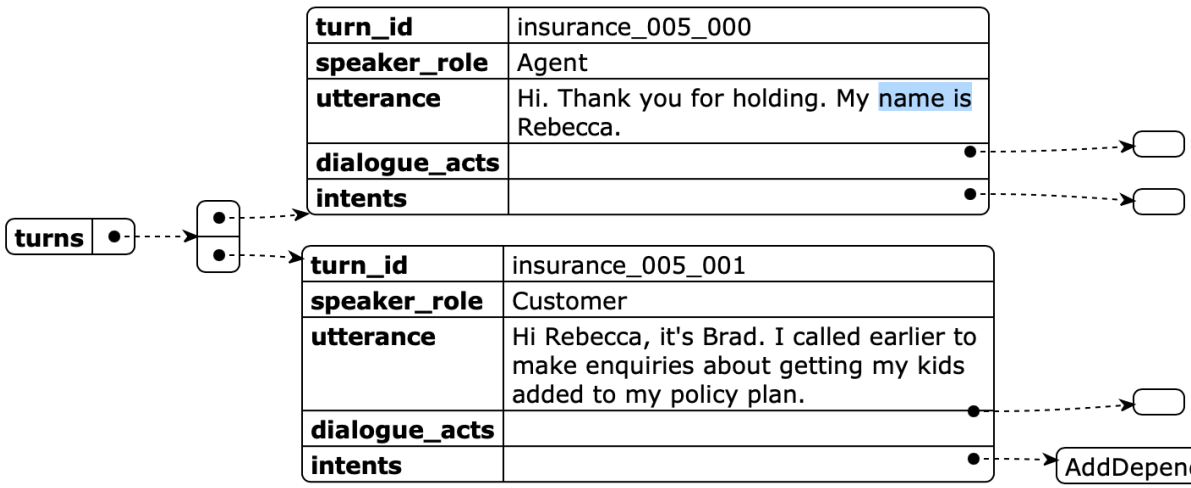
sccl

SCCL(supporting clustering with contrastive learning) is a deep clustering algorithm which can learning more adequate representation of short context. By using Contrastive Learning, same samples will be pull together while different ones will be pushing apart.



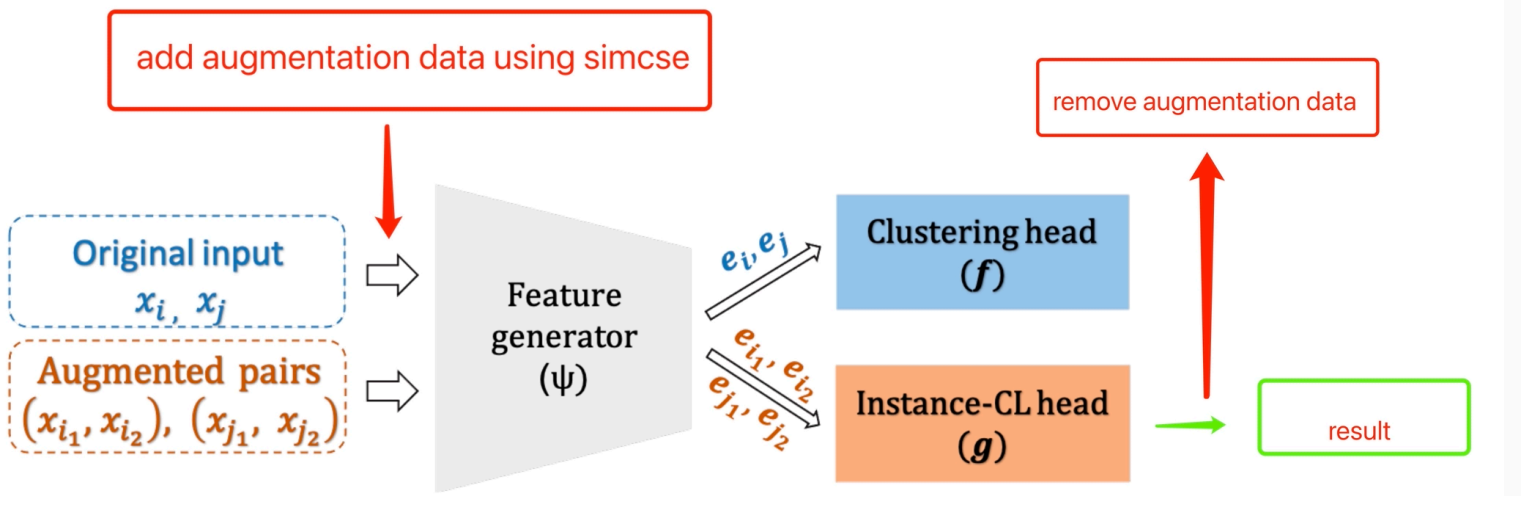
filtering high quality utterances with statistic rule or hdbscan

In open intent induction, we will not have acess to the ground-truth utterances with intents, so need to filter out noise utterances. A tradition way is to use clustering algorithms, like hdbscan, but in this task, we find a more effective way to select high quality utterances, that is statistic rules. Specificlly, we count the highest frequency bi-gram phrase before the utterances in development dataset and find it domain-agnostic, thus we migrate the rule to test dataset and obtain high quality utterances.
For example, if "insurance$_{005001}$" is a high quality utterance, "name is" is a bi-gram phrase before it.

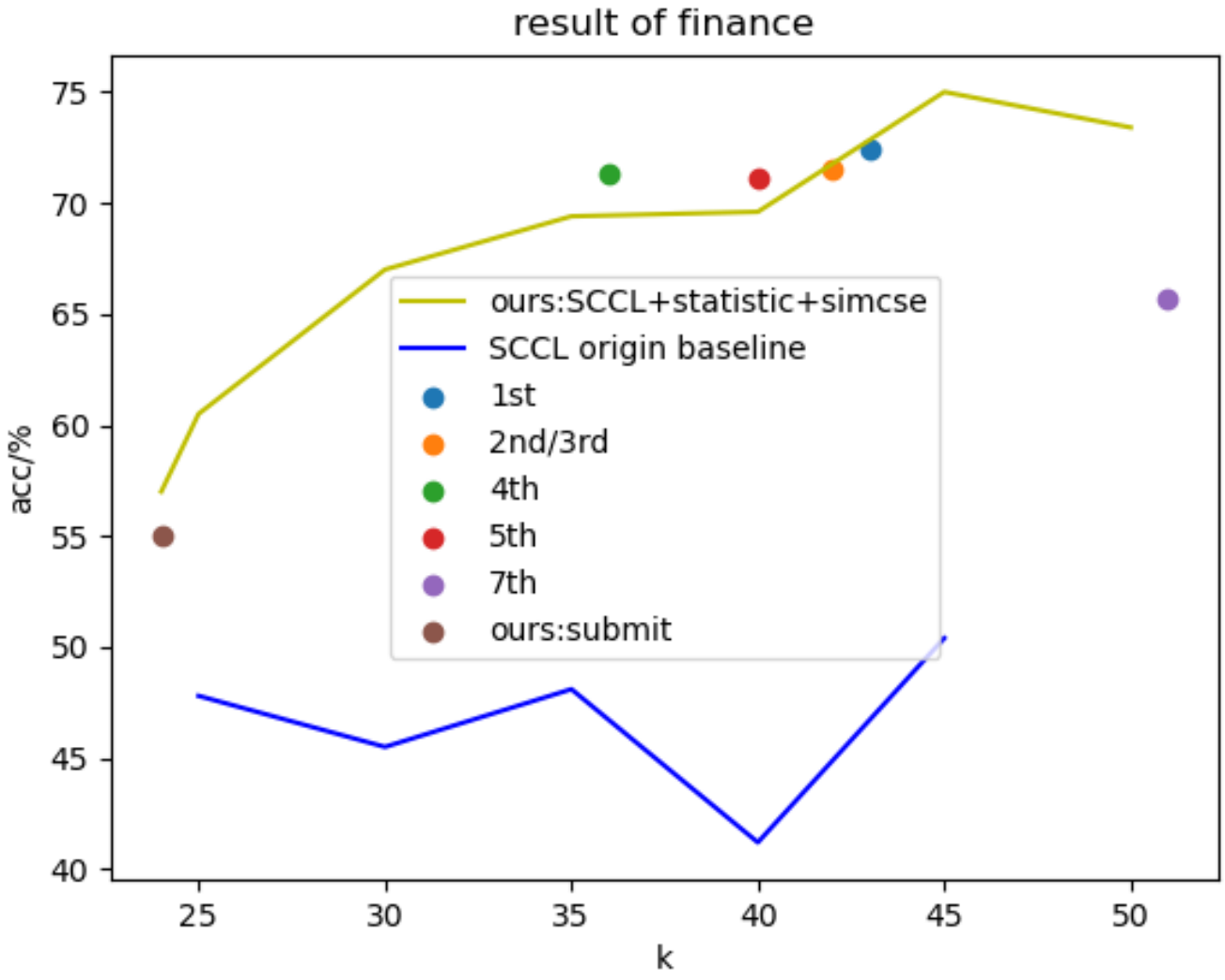


data augmentation using out-domain data

After getting high quality utterances for sccl(original input in figure), besides drop-out augmentation in sccl, we can augment the original input by simcse using out-domain data(which has similar utterances in original input). Then we can train sccl normally. After getting clustering results, we have to remove the augmentation data added in previous step.



results


result of finance
acc/%
k
75
70
65
60
55
50
45
40
25
30
35
40
45
50
ours:SCCL+statistic+simcse
SCCL origin baseline
1st
2nd/3rd
4th
5th
7th
ours:submit