draft for NLP project

yansong

February 2020

Abstract

Since I am terrible at deciding specific domain of study in NLP, instead I would propose several interesting model setting on NLP task from which we can select and apply on NLP dataet we want to carry out research on.

1 Deep Reinforcement learning model in NLP

[2]

There has been two ways to integrate these concept. One is using RL model to handle NLP dataset and the other one is assisting RL decision process with NLP data. Recent advances in representation learning for language make it possible to build models that acquire world knowledge from text corpora and integrate this knowledge into downstream decision making problems. The former is aiming at formulate the NLP problem to a RL task, the latter normally transfer knowledge from natural language corpora to RL tasks, as well as between tasks, consequently unlocking RL for more diverse and real-world tasks.

Recent advances in representation learning for language make it possible to build models that acquire world knowledge from text corpora and integrate this knowledge into downstream decision making problems.

To realize the potential of language in RL, it was advocated for more research into learning from unstructured or descriptive language corpora, with a greater use of NLP tools like pre-trained language models. Such research also requires development of more challenging environments that reflect the semantics and diversity of the real world.

2 Active learning

[1]

Machine learning in NLP require large amount of training data, which are hard to get due to costly labelling procedure. Researches related to semi-supervised learning and unsupervised learning were done to obviate the need of labeled data to a certain extent.

Other problems of labeled data are, for example, the data similar to what the learner has already seen are not as useful as new data. Correspondingly, active learning is the task of reducing the amount of labeled data required to learn the concept by querying the user for labels for the most informative data so that the target concept is learnt with fewer data.

A classical active learning setup typically consists of a small set of labeled examples and a large set of unlabeled examples. An initial classifier is trained on the labeled examples and/or the unlabeled examples. From the pool of unlabeled examples, *selective sampling* is used to create a small subset of examples for the user to label. This iterative process of training, selective sampling and annotation is repeated until convergence.

In an NLP task, active learning often include information and feature extraction, and it can also be put in an online learning framework.

3 Neural network model

Transformer (BERT), Attention network, RNN, CNN ....

References

[1] S. Arora and S. Agarwal. Active learning for natural language processing. *Language Technologies Institute School of Computer Science Carnegie Mellon University*, 2007.

[2] J. Luketina, N. Nardelli, G. Farquhar, J. Foerster, J. Andreas, E. Grefenstette, S. Whiteson, and T. Rocktäschel. A survey of reinforcement learning informed by natural language. *arXiv preprint arXiv:1906.03926*, 2019.