Decrease the Effect of Noisy Labels in Crowdsourcing NLP tasks Using Machine Learning Methods

Summary:

The performance of machine learning methods highly depends on the quality of labels using to train classifiers. However, providing such an accurate label are costly in terms of time and money. Traditionally, these labels are provided by experts for different tasks. (e.g. image categorization, speech transcription, machine translation and etc.)

Recently Natural Language Processing (NLP) community can get the advantage of a new paradigm called crowdsourcing to provide the annotations for different domains. Basically, crowd sourcing a task is hiring on-demand labors to do the annotation tasks via web medium in a form of small units. Crowdsourcing a task is cheap and fast but not reliable in terms of quality. The online workers are anonymous and have different expertise. Therefore, the generated data is noisy.

Project Description:

The aim of the project is to try various statistical solutions like machine learning methods to reduce the effect of the noisy labels in training process. (applied to NLP tasks e.g. question classification, sentiment analysis, question answering, etc.)

Sub Tasks:

1.Crowdsource a dataset of an interest domain and collect the labels. The task consists of :

- 1.1 Design the task in Crowd Flower platform,
- 1.2 Define gold questions,
- 1.3 Analysis the collected data.

"You need basic knowledge of HTML/Java script/PHP"

2. Propose simple solutions to find the outlier annotators(noisy labels)

- 2.1 Task analysis and modeling
- 2.2 Crowd workers accuracy modeling
- 2.3 Hybrid models
- 2.4 etc.

3.Train a classifier with clean and noisy labels and compare the results.