PhD Oral Defense

Date: **7 September 2020 (Monday)** **Time:** **3:30pm**

Thesis Title

Sequence Labeling with Multiple Annotations

Ms LU Xiaolei (Supervisor: Prof Tommy Chow)

Abstract

Sequence labeling has been widely applied in Natural Language Processing and Computational Biology. Training sequence models requires large amount of sequential data with exact annotations, which is costly and laborious to produce. Crowdsourcing has received increasingly attention as it provides an efficient and cheap way to obtain large labeled sequential datasets from a group of ordinary people. In the thesis, we focus on handling sequence labeling with multiple annotations.

First, by analyzing the graph structure of existing probabilistic models for sequence labeling, we propose Semi-Markov Condition Random Fields with duration modeling (DM-SMCRFs) and apply DM-SMCRFs to keyphrase extraction. Then we consider crowd sequential annotations where the quality of label sequence relies on the expertise level of annotators in capturing internal dependencies, and propose Modeling sequential annotation for sequence labeling with crowds (SA-SLC). Different from crowd sequential annotations, partial sequence labeling assumes that the ground-truth label sequence is masked by multiple annotations. By addressing that existing disambiguation strategies for partial sequence labeling just cannot generalize well to solve the problem that there are some candidates which can be false positive or similar to the ground-truth label, we propose a novel weak disambiguation for partial sequence labeling (WD-PSL). Furthermore, structured Gaussian Processes model is developed to provide a non-parametric Bayesian approach to partial sequence labeling (SGPPSL).

In summary, we explore the information of multiple annotations from annotators' perspective and label distribution. The proposed models, which are verified by extensive comparison, effectively solve the addressed problems in sequence labeling with multiple annotations.