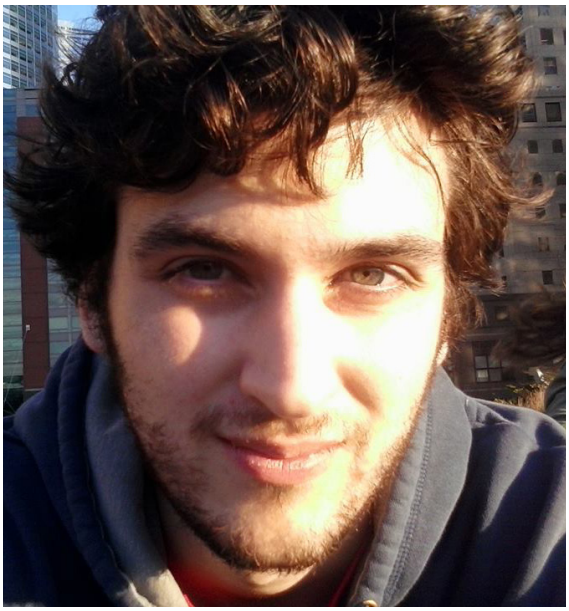
COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

DISSERTATION DEFENSE



CHARLES WELCH

Leveraging Longitudinal Data for Personalized Prediction and Word Representations

Wednesday, December 9, 2020
3:00 PM
Virtual

ABSTRACT: This thesis focuses on personalization, word representations, and longitudinal dialog. We first look at users expressions of individual preferences. In this targeted sentiment task, we find that we can improve entity extraction and sentiment classification using domain lexicons and linear term weighting. This task is important to personalization and dialog systems, as targets need to be identified in conversation and personal preferences affect how the system should react. Then we examine individuals with large amounts of personal conversational data in order to better predict what people will say. We consider extra-linguistic features that can be used to predict behavior and to predict the relationship between interlocutors. We show that these features improve over just using message content and that training on personal data leads to much better performance than training on a sample from all other users. We look not just at using personal data for these end-tasks, but also constructing personalized word representations. When we have a lot of data for an individual, we create personalized word embeddings that improve performance on language modeling and authorship attribution. When we have limited data, but we have user demographics, we can instead construct demographic word embeddings. We show that these representations improve language modeling and word association performance. When we do not have demographic information, we show that using a small amount of data from an individual, we can calculate similarity to existing users and interpolate or leverage data from these users to improve language modeling performance. Using these types of personalized word representations, we are able to provide insight into what words vary more across users and demographics. The kind of personalized representations that we introduce in this work allow for applications such as predictive typing, style transfer, and dialog systems. Importantly, they also have the potential to enable more equitable language models, with improved performance for those demographic groups that have little representation in the data.

CHAIR: Prof. Rada Mihalcea