Zusammenfassung:
Social media has produced sheer volumes of text, yet the text is short and noisy, making it hard to be harnessed by off-the-shelf NLP tools. This talk covers two data preprocessing approaches for social media text to improve its utility for NLP: (1) Text normalisation is to restore nonstandard words to their canonical forms so that the text can be better recognised and utilised by NLP tools, e.g., 2mrw ("tomorrow"), earthquick ("earthquake"); (2) Geolocation prediction disambiguates social media users’ geographical regions, i.e., it enables dividing and grouping social media text based on locations which is essential for many applications, e.g., local event detection and sentiment analysis by regions.

Kurzlebenslauf:
Bo Han is a PhD candidate at the Computing and Information System Department at the University of Melbourne. His PhD topic is to improve the utility of social media data by text preprocessing for Natural Language Processing (NLP). He has published several papers on social media text normalisation and user geolocation prediction in ACL, COLING and EMNLP. Before his PhD, Bo spent one year internship working on social media text semantic role labelling and computer assisted English writing in Natural Language Computing (NLC) group at Microsoft Research Asia.