Computer Science Seminar

Machine Translation for All

Huda Khayrallah
Johns Hopkins University

Abstract: Machine translation uses machine learning to automatically translate text from one language to another and has the potential to reduce language barriers. Recent improvements in machine translation have made it more widely-usable, partly due to deep neural network approaches. Howeverlike most deep learning algorithmsneural machine translation is sensitive to the quantity and quality of training data, and therefore produces poor translations for some languages and styles of text. Machine translation training data typically comes in the form of parallel textsentences translated between the two languages of interest. Limited quantities of parallel text are available for most language pairs, leading to a low-resource problem. Even when training data is available in the desired language pair, it is frequently formal textleading to a domain mismatch when models are used to translate a different type of data, such as social media or medical text. Neural machine translation currently performs poorly in low-resource and domain mismatch settings; my work aims to overcome these limitations, and make machine translation a useful tool for all users. ¡br¿¡br¿ In this talk, I will discuss a method for improving translation in low resource settingsSimulated Multiple Reference Training (SMRT; Khayrallah et al., 2020)which uses a paraphraser to simulate training on all possible translations per sentence. I will also discuss work on improving domain adaptation (Khayrallah et al., 2018), and work on analyzing the effect of noisy training data (Khayrallah and Koehn, 2018).

Monday, February 15, 2021, 10:00 am
https://emory.zoom.us/j/92558356951

Computer Science
Emory University