

Using Natural Language Processing to Understand Us

Sabrina Tiun

Asian language processing (ASLAN) Lab, CAIT, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor Malaysia

Abstract

The latest popular approach in natural language processing is the so-called word embedding (WE). The word embedding (WE) model is a neural network based on the distributional semantic model. The distributional hypothesis states that semantically similar words tend to have similar contextual distributions. In the WE context, if two words have similar vectors, then they have the same distribution. An application of WE based on periodical is called temporal word embedding or dynamic word embedding (DWE). We have explored the use of WE and DWE to mine lifestyle, sentiment and evolution of trends and policy. In the work of WE on Malaysian Twitter corpus, we explore the possibility of viewing Malaysia's lifestyle on where they spend most of their time for social meetings and analyse the sentiment of a public figure. The TWE model applications fall under two circumstances: For a TWE with a time-series corpus of a more extended period (or also known as a diachronic corpus), usually the related research concerns on language topics, i.e., the study on the meaning of words over time. A time-series corpus with a shorter period (called temporal corpus) is more on mining texts for culture semantic shift, i.e., detecting event for an actionable purpose. In our work on DWE, we explore how DWE can be used to see specific issues differently at different kinds of periods. Based on a Malaysian Hansard Parliamentary DWE model, the concept of 'security during the early period is more related to external (international) affairs than the recent period, which is more on local (domestic) affairs. We are currently trying to explore the sentiment of Malaysian regarding the Covid-19 pandemic impact on various aspects, such as economy, education, etc. The work presented has been funded by the Ministry of Higher Education Malaysia under research code: FRGS/1/2020/ICT02/UKM/02/1

Biography

Sabrina Tiun is a senior lecturer at the Faculty of Information Science and Technology (FTSM) in Universiti Kebangsaan Malaysia (UKM), Malaysia. Her research interests range from natural language processing, computational linguistic, and speech processing to social science computing. She teaches several courses in natural language processing and data science at both undergraduate and postgraduate levels.

She also actively and successfully supervises postgraduate students for master and PhD level in natural language processing and speech processing. Currently, she is the head lab of Asian language processing (ASLAN) lab and the teaching and learning coordinator for the Centre of Artificial Intelligence Technology at FTSM.