Title: Populating a Knowledge Graph of Singapore Pioneers: Information Extraction from Biographies in Singapore Infopedia Using NLP (SpaCy) and Transformer/BERT modeling

Christopher S.G. Khoo and Siam-Gek Ng

Wee Kim Wee School of Communication & Information

Nanyang Technological University, Singapore

This paper reports on our effort in populating SingPioneers.sg -- a knowledge graph of personalities significant in the history and development of modern Singapore, based on more than 400 biographies in the Singapore Infopedia, an electronic encyclopedia of the Singapore National Library Board. A prototype knowledge graph application has been implemented at https://SingPioneers.sg using a Neo4j graph database management system as backend database, a Node.js+KOA Web API as middleware, and Cytoscape.js javascript library to implement a graph visualization Web interface. The prototype knowledge graph was handcrafted from manual analysis of the biographies of 20 selected SingPioneers.

In the second phase of the project reported in this paper, we develop information extraction methods to extract relational information from more than 400 biographies to complete the knowledge graph. We use a combination of NLP and transformer/BERT approaches. The NLP approach makes use of the SpaCy python package to extract entities and noun phrases. Handcrafted linguistic patterns comprising a combination of lexical tokens, part-of-speech tags, entity types and dependency relations are used to extract relational information from the biography texts. Subsequently, the text + extracted information will be used to train transformer/BERT models to improve the recall and precision. Comparison of the two approaches and lessons learned will be presented.