MINISTRY OF BUSINESS,
INNOVATION & EMPLOYMENT
HĪKINA WHAKATUTUKI

Summary of Natural Language Processing Discussion

NZ-SG Joint R&D Data Science workshop
30 & 31 October 2019
NUSS Kent Ridge Guild House

SINGAPORE DATA SCIENCE CONSORTIUM
NZ-SG JOINT R&D DATA SCIENCE WORKSHOP
30 & 31 October 2019
NUSS Kent Ridge Guild House

Participants

Te Taka Keegan	tetaka.keegan@waikato.ac.nz	University of Waikato
Tahu Kukutai	tahu.kukutai@waikato.ac.nz	
Albert Bifet	abifet@waikato.ac.nz	
Finlay Thompson	finlay@dragonfly.co.nz	Dragonfly Data Science
Titima Suthiwan	clsts@nus.edu.sg	National University of Singapore
Daniel Chan	daniel.chan@nus.edu.sg	
Sasiwimol Klayklueng	clsnm@nus.edu.sg	
Tan Kian Lee	tankl@comp.nus.edu.sg	
Ng See Kiong	seekiong@nus.edu.sg	
David Lo	davidlo@smu.edu.sg	Singapore Management University
DAI Bing Tian	btdai@smu.edu.sg	
Zhengkui Wang	Zhengkui.Wang@singaporetech.edu.sg	Singapore Institute of

		Technology
Jung-jae Kim	jjkim@i2r.a-star.edu.sg	Institute for Infocomm Research, A*STAR

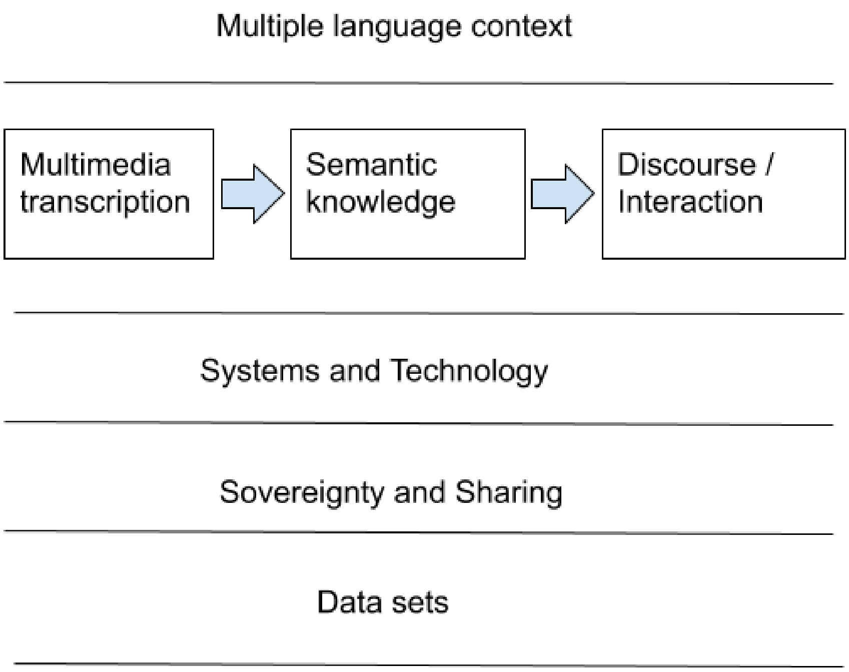
Wednesday 30 October

The session began with introductions around the table, followed by general discussion and sharing about the mutual interests in languages in New Zealand (Māori languages) and Singapore (SEA languages, Singlish and local dialects), and their roles in culture and heritage.

The objectives of the brainstorming session were then discussed and everyone invited to brainstorm ideas written on post-it notes for the whiteboard.

The post-it notes were organised into various clusters by See-Kiong, followed by group discussions to finetune the clusters and classify each cluster with an appropriate research topic. This led to the overall structure of the various NLP research substreams as depicted in the figure below:

OVERALL STRUCTURE OF NLP RESEARCH



SUBSTREAMS

The brainstormed ideas for data science/AI research in each identified NLP substream included:

DATASETS

- Reviews of books and titles
- Social Media data
- Product review stream analysis (digital products, rapidly changing, stream of reviews from large user base)
- Corpus / concordance to learn appropriate collation
- AI to determine the accuracy of language data

SOVEREIGNTY AND SHARING

- Privacy-preservation in text data
- Federated computing in NLP systems for inference sharing

SYSTEMS & TECHNOLOGY

- Transcription algorithms on the edge
- ML / DS on the edge
- Management of large quantity of data
- Efficient query processing
- Green data science to reduce CO_2 consumption
- Duplicate elimination (clustering of similar text)
- Data cleaning and processing techniques

MULTIMEDIA TRANSCRIPTION

- Pronunciation assist for learners, NZ Waikato Te Reo Maori
- Speech recognition of learner language and error analysis/correction (no data at this stage?)
- Transcription for low resource languages
- Multimedia data, transcription from video

SEMANTICS & KNOWLEDGE

- Semantic / knowledge analysis of old written texts, e.g. NZ land court records
- Knowledge graph construction and use for task-specific chatbot
- Text mining to identify knowledge structures
- AI to identify knowledge structure and meaning
- How to detect and extract alternative knowledge system. Semantic data from multiple and fragmented data sets
- Unlocking and understanding the value of heritage data sets
- How to build small dialectical language and cultural context data into AI tools that predict …. E.g. kinship
- Q & A from text and knowledge graph
- NLP for vulnerability detection and patching (cybersecurity)

DISCOURSE & INTERACTIONS

- Use data science to facilitate learning of minority languages
- Language learning or teaching with chatbots
- Technology enhanced language learning, detection of tones

MACHINE LEARNING IN MULTI-LANGUAGE CONTEXT

- Transfer learning across languages

- Transfer learning between different language datasets
- Cross language learning
- Remodeling NLP techniques for other languages / dialects
- Federated machine learning models (multi-language text data, semantic analysis)
- How to automate identification and analysis of heritage / minority multiple language knowledge
- Language identification in multilingual text datasets

Thursday 31 October

The second day was focused on generating three example project ideas.

First, the group brainstormed on various application domains that are of mutual interest to SG and NZ:

SUGGESTED APPLICATION DOMAINS

- Education (e.g. mobile learning)
- Cultural / Heritage preservation and revitalisation
- Intercultural understanding
- Tourism
- Product / customer analysis
- Translation

It was noted that some data sets and applications are more commercially-focused, while some more cultural heritage. Both commercial and cultural aspects are of interest to SG and NZ.

SUGGESTED PROJECT IDEAS

- Towards a global NLP platform and beyond: NZ-SG federated learning platform for community learning..
    - A federated learning platform that supports cross-geographical research in NLP (and other research topics of interest such as health and 3D spatial data) between Singapore and NZ, and beyond.
- Multi-modal, Multi-lingual and Multi-cultural Language Processing and Understanding
    - Audio + Video to text transcription. Dealing with gestures, and feeding into database and query language for Q&A tool. Provide tool for capturing institutional memory (including corporate settings -- retirees),

oral histories, educational videos. Sentiment analysis beyond word clustering into syntax and grammar.

  - Word level language identification. E.g. singlish, x-lish, other heritage languages in Singapore, and te reo Māori / NZ english. Involves a mixed language and multi-language model. Code switching. Technology as a lever to counter drive towards linguistic (and epistemic) homogeneity.
  - Semantic understanding of multilingual and fragmented documents. E.g. Land court documents, other archival fragments from multiple sources, and land tenure. Incorporation with traditional/contextual knowledge conveyed through elders.
  - Cross-cultural knowledge graph discovery. Common understanding of “success”, “family”, “happiness”. A data-driven approach to comparative sociology. Temporal evolution of social values.
- NLP in New Media
  - Text/video data stream analysis. Early detection of non-binary or negative sentiment, event detection, online learning of new words, temporal evolution of knowedge graph, fake news detection, social polarisation. False binaries. Simultaneous speakers.
  - An example detailed project on social media data stream processing has been prepared David, Albert, and Zhengkui in Appendix A.

Appendix A

Example Project:
Making Sense of Multi-Lingual Multi-Cultural Text Data Streams in New Media SG + NZ Scale

In NZ and Singapore, both residents and non-residents (eg tourists) generate a large stream of natural language and multimedia contents (e.g., tweets, images, videos) in various social media platforms (e.g., Twitter, Instagram, Facebook). An efficient and effective solution is needed to process this rich source of information for various **purposes** that matter for both New Zealand and Singapore, such as:

(Bad / Emerging) Event Detection

In the age of social media, it may be possible to detect (bad or emerging) events early as they appear in the data stream. For example, a particular accident may happen and decision makers (e.g., ambulance, firemen, etc.) may want to take actions to respond to the event promptly. A particular problem may occur for a tourist attraction and we want to detect it early before the issue becomes viral.

Collective Heritage Preservation and Revitalization

Heritage are a nation's unique wealth which defines its identity. Can data science and AI technologies help in the preservation and revitalization of our heritage? As a member of a community, one may want to share the little piece of heritage that he/she has that can then be digitalized, preserved, and stored in a query-able format. This can be done by posting relevant cultural/heritage items, e.g., stories, cook books, songs, interviews, etc. in the form of videos, images, etc. that can be tagged with particular keywords for collection and curation.

Participatory Decision Making

Residents and tourists may have in mind some thoughts about e.g., a certain policy, new "features" of a particular tourist attraction, etc. They may share their "wishes" in social media through text, images, or even videos. There is a need for effective tools that can enable one to gather these inputs from people to help in participatory decision making to enable effective decisions that involve all "stakeholders".

Racial Harmony

Small issues can become large if it is not handled properly (e.g., recent incidents in Papua, Indonesia). Detecting such issues before they blow up into big problems (e.g., racial riot in Papua that cause many casualties) can aid decision makers to take preventive measures.

https://www.theguardian.com/world/2019/sep/28/i-feel-like-im-dying-west-papua-witnesses-unrest-indonesia-police

De-"polarizing" Echo Chambers

Social media users form "echo chambers". They can only "see" contents generated by people who are similar to them. In a way, they only look at "trees" and miss the "forest". It would be helpful to design solutions that can enable social media users to get a "summary" of various viewpoints towards a particular issue in real time. For many issues (e.g., Brexit), often there are many viewpoints and not only one. Some news media today may often only cover "one side of a coin".

https://www.channelnewsasia.com/news/singapore/in-a-world-of-echo-chambers-that-polarise-stories-help-build-bri-7606668

Detecting Fake News

Fake news are often propagated on social media. Singapore just passes a law about fake news. It would be interesting to detect such fake news by comparing different “streams” of data (e.g., streams from trusted news agencies, well-known figures vs. streams from dubious / questionable individuals).

https://www.straitstimes.com/politics/fake-news-law-to-come-into-effect-oct-2

Solving the above require a number of **datasets**:

- SMU has a dataset of tweets generated by social media users in Singapore. This tweet can be used in SMU but not propagated to third parties, due to Twitter data sharing policy.
- According to his website, Keegan from U. of Waikato has also analyzed tweets written in Maori language.

Solving the above require a number of **technical challenges** which would require the development of data science and AI to address:

- Unlike the other countries, for Singapore and NZ, social media data (e.g., tweets) can be written in multiple languages. In Singapore, tweets can be in English, Chinese, Malay, and Tamil. In New Zealand, tweets can be in Maori and English. Tourists visiting the two countries can write tweets in many different languages.
- Data can consist of text in different languages. Detecting which language parts of a tweet / video belong to is not an easy task.
- Federated learning to allow sharing of models between Singapore and New Zealand without sharing the raw data. Note that for tweets sharing of raw data is prohibited (due to Twitter data sharing policy). Federated learning is more difficult for data stream setting due to potential adversarial setting: (1) adversary can inject new data (one can inject data by writing Tweets about Singapore or New Zealand that may influence the learning of models); (2) adversary can “diff” models that are changing in real time as more tweets are received / processed in the stream.
- Detection of new concepts / words and dealing with concept drift.
- Efficient processing of millions of streams of data coming from various social media accounts.
- Processing of contents in various forms: images, text, videos.
- Dealing with non binary sentiments (not only +ve vs. -ve but the whole spectrum of sentiments).
- Online and continuous learning via topic modelling or deep learning.
- Dealing with bias and diversity (representing everyone in the population and not letting the majority to silence the minority).
- Detection of important “features” expressed in many ways in various tweets.
- Detection of emerging events early before it went viral.
- Question answering and chat bot generation using streams of text, videos, and images as knowledge base.