Do the mahi, get the tweets!

David Trye, Department of Computer Science
Supervised by Te Taka Keegan, Andreea Calude & Felipe Bravo-Marquez

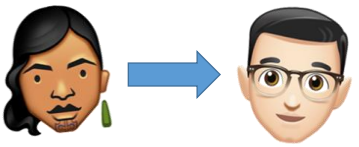
THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

waikato.ac.nz WHERE THE WORLD IS GOING

Setting the Scene (1)

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

- What are loanwords?
  - Words that are borrowed from another language
  - Arise in situations of **language contact**
- Māori loanwords have trickled into New Zealand English (NZE)
  - Two main "waves" of borrowing (Macalister, 2006)
    - Colonisation period: flora & fauna terms
    - (Ongoing) decolonisation period: social & material cultural terms
  - Used for various functions
    - To fill semantic gaps, signal solidarity, economy of expression, etc.
  - Direction of lexical transfer highly unusual

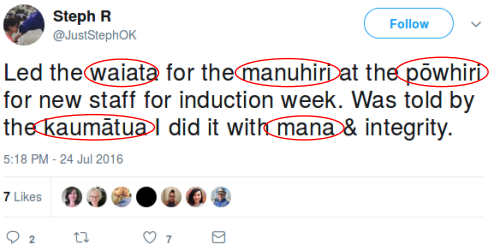


Donor language: Māori

Receiver language: English

Setting the Scene (2)

- Māori loanword use is **highly skewed**, by both topic and speaker/writer (gender & ethnicity)



- Loanword use is **increasing**
- Some loanwords "do better" than others
  - e.g. shorter words, core rather than cultural terms

Research Aims

1. To build a **corpus** of NZE tweets containing Māori loanwords
   - *Māori Loanword Twitter* (MLT) Corpus
   - Needs to be large, clean and balanced
   - Twitter data is cheap but **noisy**!

Research Aims

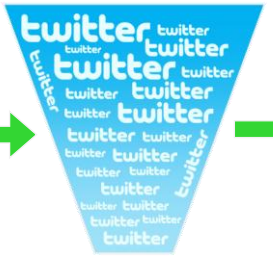
2. To **analyse** how Māori loanwords are used in the corpus
   - Surprising lack of research into how loanwords are used on social media
     - Many other genres studied
     - Twitter provides different kind of data
       - Formal & informal
       - Not edited
       - Creative
       - Single-authored
       - Normative & non-normative
   - What can social media tell us about Māori loanwords?

Building the MLT Corpus (1)

116 Loanwords
"query words"

Aotearoa
Aroha
Atua
Awa
⋮
Whero

8 million Tweets
(2007-2018)



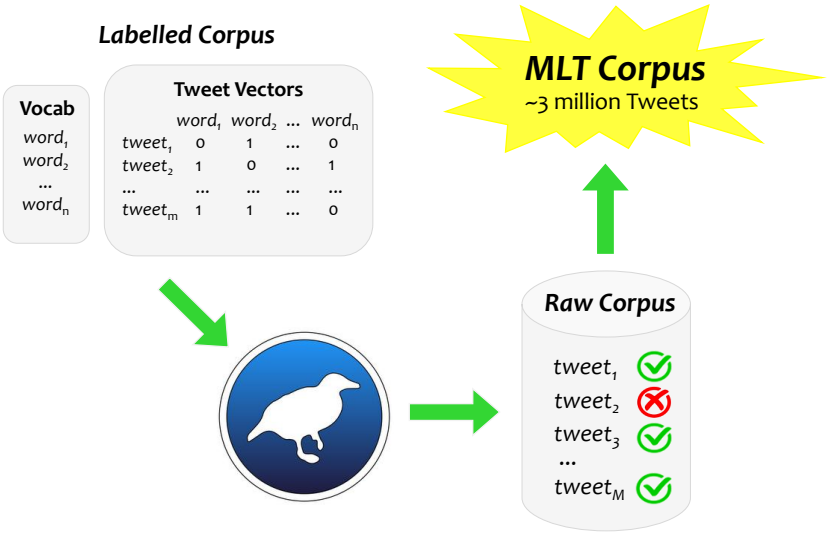
4,600 Labelled Tweets
40 tweets per query word

Proud to be a ***kiwi*** ✓
Love my crazy ***whanau*** ✓
Moana is my fav Princess ✗
haka ne kuma fa you say ✗

Raw Corpus
77 'best' loanwords
~4.5 million tweets

Labelled Corpus
77 'best' loanwords
3,600 *labelled* tweets

Building the MLT Corpus (2)



To Summarise: Training our Classifier

- Goal
  - Eliminate **noise** in MLT Corpus
    - Loanwords conflated with homographs, proper nouns, misspellings, foreign languages, etc.
- Solution
  - Build a machine learning model to automatically detect whether a tweet is **relevant** (i.e. used in a NZE context) or irrelevant
  - Split data into *training data* for building the model and *test data* for evaluating it (the latter not seen during training)
  - *Probabilistic* binary classification:

$$f(x) = \begin{cases} irrelevant & \text{if } x < 0.5 \\ relevant & \text{if } x \geq 0.5 \end{cases}$$

- Supervised learning approach
  - We use **labelled data** to predict class labels for new (unseen) instances

Machine Learning Input and Output

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Training Data (Input)

id	username	timestamp	query word	text	relevance (gold label)
7573693 4364248 0640	JustStephOK	2016-07-25 12:18	***waiata***	Led the ***waiata*** for the manuhiri at the pōwhiri for new staff for induction week. Was told by the kaumātua I did it with mana & integrity.	**relevant**

Target Data (Output)

id	username	timestamp	query word	text	f(x)
8095892 4403756 6460	KUOI_DJ	2016-12-16 15:41	***waiata***	Split Enz—History Never Repeats— ***Waiata***	0.078 (**irrelevant**)

Model Evaluation (1)

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

- Complex classification problem
  - Class label depends on both *context* and *query word*
  - Domain overlap
    - Irrelevant context for one query word might be relevant for another
    - "singing" and *kiwi* (irrelevant) vs. "singing" and *waiata* (relevant)
  - Created own independent **stratified samples**
    - Instead of using (randomised) cross-validation
    - To maintain distribution of relevant/irrelevant tweets for each query word, as seen in the labelled corpus
    - 80/20 split for training and test data

Model Evaluation (2)

- Can't rely on **observed accuracy** when class distribution is skewed
  - 2/3 relevant tweets (majority class)
  - 1/3 irrelevant tweets (minority class)
- Instead, we chose to evaluate our models using:
  - Kappa
    - Were correct classifications obtained simply by chance?
    - Ranges from 0 to 1 (best)
  - AUC
    - Area under the ROC curve
    - Calculated by plotting true positive rate (TPR) against false positive rate (FPR) at various thresholds
    - Ranges from 0.5 to 1 (best)
  - Weighted average F-Score
    - Combines precision and recall
    - Ranges from 0 to 1 (best)

Classification Results on Test Set

	Word *n*-grams	AUC	Kappa	F-Score
Naive Bayes Multinomial	1	**0.872**	**0.570**	**0.817**
Logistic Regression	1	0.863	0.534	0.801
	1, 2	0.868	**0.570**	0.816
	1, 2, 3	0.869	0.560	0.811
	1, 2, 3, 4	0.869	0.563	0.813
	1, 2, 3, 4, 5	0.869	0.556	0.810

Corpus Statistics

	Tokens (words)	Tweets	Tweeters (authors)
Raw Corpus	70,964,941	4,559,105	1,839,707
Labelled Corpus	49,477	2,495	1,866
Processed Corpus	47,547,878	**2,955,450**	1,256,317

Preliminary Findings

- **Code-Switching**
  - Alternating between English and Māori in same tweet
  - Clauses or sentences (rather than individual words)

Preliminary Findings

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

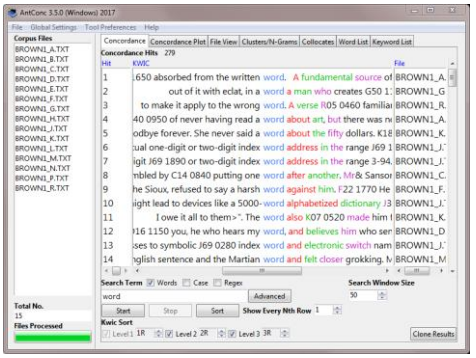
- **Hybrid Hashtags**
  - Hashtags that contain lexical items from two or more languages (in our case, English and Māori)
    - #growing-up-**kiwi**
    - #**kai**-to-put-in-my-fridge
    - #trans-**whanau** ...

  - We intend to analyse their syntactic structure, discourse function and frequency & use over time

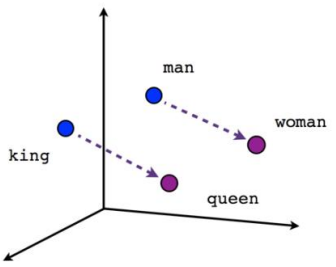
Word Embeddings (Mikolov et al. 2013)

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Distributional Hypothesis:
"You shall know a word by the company it keeps"
(John Firth, 1957)



Word2Vec



Similar words receive similar vectors, and are thus situated closer together in *n*-dimensional space

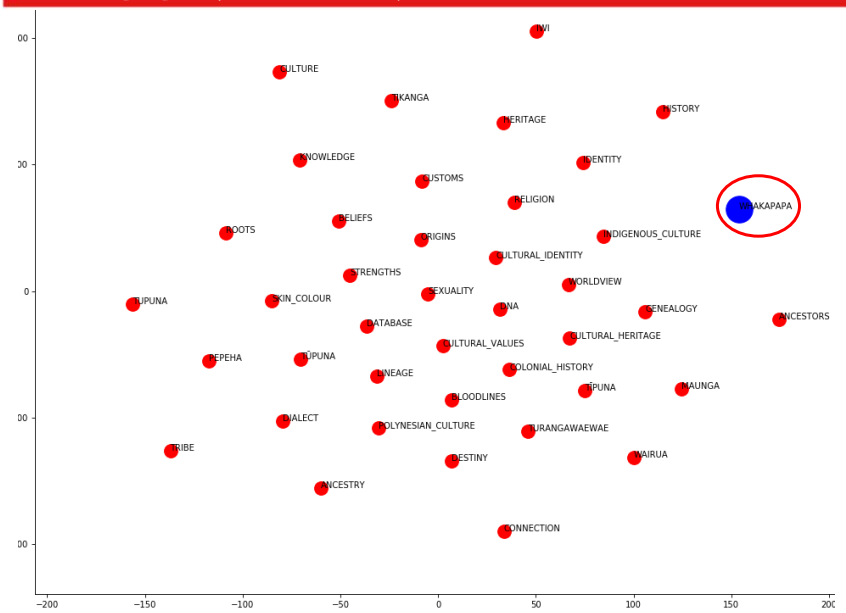
Image source: http://www.laurenceanthony.net/software/antconc/screenshots/AntConc_1.png

Training Word Embeddings

- Hyper-parameters optimised
  - By minimising **median ranking** of a list of "gold pairs" of Māori words and their *distinct* English counterparts
  - Window size (*w*) = number of words to consider on either side of target word
  - Vector size (*v*) = number of dimensions
- Plot graphs for each loanword
  - Using t-SNE to project embeddings downwards into 2D space

Māori	English
aotearoa	new_zealand
morena	good_morning
aroha	love
haka	dance
kura	school
mauri	life
whanau	family
wahine	woman
iwi	tribe
pakeha	european
tautoko	support
matua	parent
tangata	people
taringa	ear
taonga	treasure
korero	talk
whakapapa	genealogy
mokai	slave
reo	language

Word Embedding Plot for *Whakapapa* (w=5, v=100)



Conclusions

- First purpose-built, large-scale corpus of NZE tweets
  - **kiwiwords.cms.waikato.ac.nz**
- New methodology for filtering out irrelevant tweets
  - Using supervised machine learning
  - *Lots* of pre-processing needed to obtain data suitable for linguistic analysis
  - **8 million** tweets reduced to **3 million**
- Code-switching and hybrid hashtags pose interesting research questions and merit further study
- Word embeddings can provide valuable insights into understanding the semantic make-up of loanwords

Questions

- Thanks for listening!

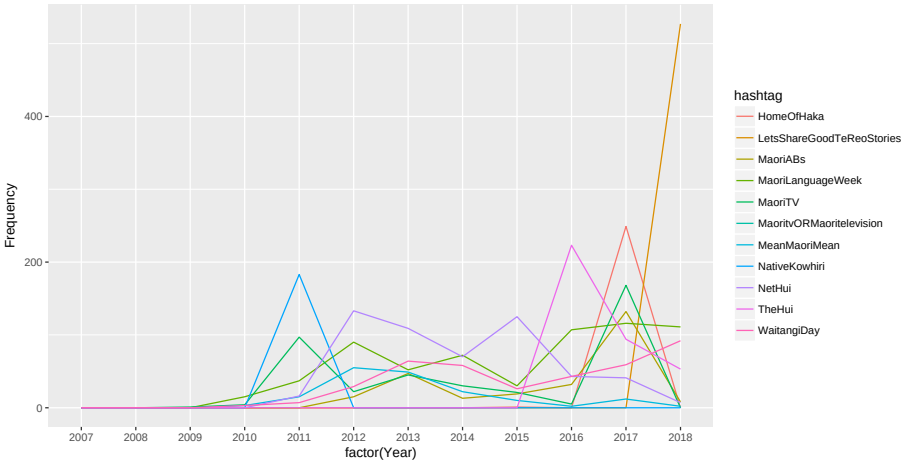
Future Work: Expanding the Corpus

- Lexicon classifier to automatically detect **Māori words/phrases** in MLT corpus
  - Character n-grams instead of word n-grams
  - Using English and Māori wordlists as training data (and under-sampling English)
  - Model classifies each word as English or Māori with probability estimate
- Use output to identify most frequent loanwords in corpus
  - Can then our supplement our original list of query words
  - Collect additional tweets -> increase size of corpus
  - More data (and target words) for training word embeddings
  - Repeat (iterative process)
- Could also use this classifier to extract all tweets that contain code-switching
  - e.g. at least four adjacent Māori words

Pre-processing

- Ensured tweets (mostly) written in **English**
- **Lower-cased** tweets & query words
- Retained **stop words**
- For **macron** words, searched with and without macrons
  māori and **maori**
- For **phrases**, searched with and without space
  kai moana and **kaimoana**
- Removed **retweets**
- Removed tweets containing **URLs**
- Removed tweets where query word part of **username** or **mention**
  @happy_**kiwi**
- Removed **short tweets** (<5 words)
- Removed **duplicate** tweets (with same ids)
  Containing multiple query words
- Removed **near-identical** tweets (with different ids)
  Differ only by punctuation, emoticons and/or *@user* mentions

Hybrid Hashtags: Use over Time



Why not Deep Learning?

- Corpus **not large enough** to see significant improvement
- **Advantages** of using probabilistic models
  - Representation more intuitive
  - Easier to interpret
  - Incorporates constraints and uncertainty