Do the mahi, get the tweets!
David Trye, Department of Computer Science
Supervised by Te Taka Keegan, Andreea Calude & Felipe Bravo-Marquez

Setting the Scene (1)

• 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

Raw Corpus
77 ‘best’ loanwords
~4.5 million tweets

Labelled Corpus
77 ‘best’ loanwords
3,600 labelled tweets
Building the MLT Corpus (2)

Labelled Corpus

Tweet Vectors

<table>
<thead>
<tr>
<th>word_1</th>
<th>word_2</th>
<th>...</th>
<th>word_n</th>
</tr>
</thead>
<tbody>
<tr>
<td>tweet_1</td>
<td>0</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>tweet_2</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>tweet_m</td>
<td>1</td>
<td>1</td>
<td>...</td>
</tr>
</tbody>
</table>

Vocab

word_1

word_2

...

word_n

MLT Corpus

~3 million Tweets

Raw Corpus

| tweet_1 | ✔️    |
|         |       |
| tweet_2 | ✗    |
|         |       |
| tweet_3 | ✔️    |
|         |       |
|         | ✔️    |

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:

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

• Supervised learning approach
  • We use labelled data to predict class labels for new (unseen) instances
Model Evaluation (1)

- 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

<table>
<thead>
<tr>
<th>Word n-grams</th>
<th>AUC</th>
<th>Kappa</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Bayes Multinomial</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.872</td>
<td>0.570</td>
<td>0.817</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.863</td>
<td>0.534</td>
<td>0.801</td>
</tr>
<tr>
<td>1, 2</td>
<td>0.868</td>
<td>0.570</td>
<td>0.816</td>
</tr>
<tr>
<td>1, 2, 3</td>
<td>0.869</td>
<td>0.560</td>
<td>0.811</td>
</tr>
<tr>
<td>1, 2, 3, 4</td>
<td>0.869</td>
<td>0.563</td>
<td>0.813</td>
</tr>
<tr>
<td>1, 2, 3, 4, 5</td>
<td>0.869</td>
<td>0.556</td>
<td>0.810</td>
</tr>
</tbody>
</table>
## Corpus Statistics

<table>
<thead>
<tr>
<th></th>
<th>Tokens (words)</th>
<th>Tweets</th>
<th>Tweeters (authors)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Raw Corpus</strong></td>
<td>70,964,941</td>
<td>4,559,105</td>
<td>1,839,707</td>
</tr>
<tr>
<td><strong>Labelled Corpus</strong></td>
<td>49,477</td>
<td>2,495</td>
<td>1,866</td>
</tr>
<tr>
<td><strong>Processed Corpus</strong></td>
<td>47,547,878</td>
<td>2,955,450</td>
<td>1,256,317</td>
</tr>
</tbody>
</table>

## Preliminary Findings

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

---

Tweet example:

```
Muzzagain
@muzzagain

> Heh! He porangi toku ngeru. especially at 5 in the morning!!

> Ata marie e hoa ma. I am well thank you.😊
```

6:46 AM · Sep 16, 2017 · Twitter Web Client
Preliminary Findings

• 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)

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

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

<table>
<thead>
<tr>
<th>Maori</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>aroaroa</td>
<td>new_zealand</td>
</tr>
<tr>
<td>morena</td>
<td>good_morning</td>
</tr>
<tr>
<td>aroha</td>
<td>love</td>
</tr>
<tr>
<td>haka</td>
<td>dance</td>
</tr>
<tr>
<td>kura</td>
<td>school</td>
</tr>
<tr>
<td>maunui</td>
<td>life</td>
</tr>
<tr>
<td>whanau</td>
<td>family</td>
</tr>
<tr>
<td>wahine</td>
<td>woman</td>
</tr>
<tr>
<td>kāi</td>
<td>tribe</td>
</tr>
<tr>
<td>peiha</td>
<td>europe</td>
</tr>
<tr>
<td>tautoko</td>
<td>support</td>
</tr>
<tr>
<td>matua</td>
<td>parent</td>
</tr>
<tr>
<td>tangata</td>
<td>people</td>
</tr>
<tr>
<td>taringa</td>
<td>ear</td>
</tr>
<tr>
<td>taonga</td>
<td>treasure</td>
</tr>
<tr>
<td>kōrero</td>
<td>talk</td>
</tr>
<tr>
<td>whakapapa</td>
<td>genealogy</td>
</tr>
<tr>
<td>mokai</td>
<td>slave</td>
</tr>
<tr>
<td>reo</td>
<td>language</td>
</tr>
</tbody>
</table>

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 undersampling English)
  - Model classifies each word as English or Māori with probability estimate
- Use output to identify most frequent loanwords in corpus
  - Can then 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