Method

Collect Tweets

The indigenous language of New Zealand is Māori, spoken by roughly 4% of the New Zealand population. Historically, Māori served as an Austronesian language which constitutes the last stop on the “island-hopping” trail in originating in Taiwan.

Words that are borrowed from another language are called homographs. Māori loanwords are widely used in New Zealand English (NZ), for various social functions by New Zealanders within and outside of the Māori community. Motivated by the lack of linguistic resources for studying how Māori loanwords are used in social media, we present a new corpus: Māori New Zealand English Tweets (MNZET).

We collected tweets containing selected Māori words that are likely to be known by New Zealanders who do not speak Māori. Since over 90% of these words turned out to be resilient (e.g. kia is a popular gaming term), Moana is a character from a Disney movie, we manually annotated a sample of our tweets into relevant and irrelevant categories. This data was used to train machine learning models to automatically filter irrelevant tweets.

New Zealand English

One of the most salient features of New Zealand English (NZE) is the widespread use of Māori loanwords, such as kia (kia), kia (kia), and Katoa (katoa New Zealand). Below are four examples of real tweets containing a rich variety of loanwords (emphasized in blue):

(1) Sorry I thought you were kia (New Zealand).
(2) Moana is the Māori name for NZ (kia1214000001).
(3) I have been learning te reo [the language] for my new week of induction week. Was told by the Māori (kia) about their culture and what the importance of kia (kia) in the Māori culture (kia).
(4) The use of Māori words has been studied intensively over the past thirty years, offering a comprehensive insight into the evolution of one of the youngest dialects of English – New Zealand English (1-5). One aspect which is missing in this body of work is the online discourse presence of the loanwords. Almost all studies come from (collaborative) written language (highly edited, revised and scrutinised newspaper data, 4, 5-11, and fictional texts -8, or from spoken language collected in the late 1990s (7).

Loanwords are thought to arise in situations of language contact (i.e. when speakers of one language have contact with speakers of another). The language contact situation in New Zealand creates a unique case for loanwords, for these main reasons:

(1) The direction of lexical transfer is highly unusual: namely, from an endogenous, indigenous language (Māori) into a dominant lingua franca (English).
(2) Because Māori loanwords are “New Zealand’s and New Zealanders’ slang” [12] and show speakers’ consciousness, their study over the years provides a fruitful comparison of the use of loanwords across genres, contexts and time.
(3) Loanword use is an increasing trend (7, 9) but the reasons for this are still unclear, and require further investigation.

The MLT Corpus

We have devised a novel method of building a corpus of New Zealand English tweets which is both (relatively) clean and large (1.2 million tweets). The Māori Loanword Twitter Corpus (MLT Corpus) affords the study of Māori loanwords dialectically (over an eleven-year period) and discretely (by topic profile). To the best of our knowledge, this is the first large-scale corpus of New Zealand English tweets and the first collection of unfiltered social media data to analyse the use of Māori loanwords in New Zealand English.

Transform Dataset

We used the Tweet Search API to harvest 9 million English-language tweets containing one or more Māori words (query words). These tweets were created to create the Raw Corpus.

We extracted a random sample of tweets for each query word and labelled these as “relevant” or “irrelevant” depending on the context. The annotated tweets, which comprise the labelled corpus, became our training data (after removing all query words that were irrelevant more than 90% of the time).

Label Samples

Proud to be a kiwi
Love my crazy whānau
Moana is my fav princess
kia ne kuma fa you say

We converted our data into a suitable format for machine learning. This involved transforming the text into vectors based on the word-by-grams they contain.

Build Classifier

Using stratified, independent test and training sets in Weka, we experimented with various machine learning models, including Naïve Bayes Multinomial and Linear Logistic Regression (with different word n-grams). We evaluated the models with the best performance.

Deploy Model

We deployed this model on the Raw Corpus to obtain automatic predictions for the relevance of each tweet. We then removed all tweets that were classified as irrelevant (p<.05), thereby producing the Processed MLLT Corpus.

References

[1] S. C. Colada, School of General and Applied Linguistics, University of New Zealand. artneticle:colda@okaito.ac.nz
[2] David Trye, Department of Computer Science, University of Canterbury, dgt12@students.unbc.ca
[3] Andreea S. Colada, School of General and Applied Linguistics, University of New Zealand. artneticle:colda@okaito.ac.nz
[4] Felipe Bravo-Margarete, Department of Computer Science, University of Chile & NUI, Chile felipe.bravo@uc.cl
[5] Te Toka Keegdian, Department of Computer Science, University of Canterbury, tatakakeeged@canterbury.ac.nz

Step 2: Label Samples

We decided to address the "noisy" tweets in our data using supervised machine learning. Coders manually inspected a random sample of 30 tweets for each query word, and labelled each tweet as “relevant” or “irrelevant” (based on the context of use). To score the reliability of our annotators, we produced a dataset of 3,685 labelled tweets for the remaining 79 query words, which comprise the labelled Corpus. Based on the assumption that the coded samples represent the real distribution of relevant/irrelevant tweets for each query word, the 39 “noisy” query words were also removed from the Original Dataset. In this way, we created the Raw Corpus, which is approximately a 6% of the size (18 million tweets reduced to 1.6 million tweets).

Step 3: Deploy Model

We trained a classifier using the Labelled Corpus as training data, so that the resulting model could be deployed on the Raw Corpus. Our goal was to obtain automatic predictions for the relevance of each tweet in this corpus, according to probabilities given by our model.

We created test and training sets that maintain the same proportion of relevant and irrelevant tweets associated with each query word in the Labelled Corpus. We chose to include 80% (2,546) of these tweets in the training set and 20% (736) in the test set.

Using the AffectiveTweets package [14], our labelled tweets were transformed into feature vectors based on the word vectors they contain. We then trained various classification models on this transformed data in Weka. The models we tested were 1) Naïve Bayes Multinomial with unigram attributes and 2) L2-regularized logistic regression models with different word-ngram features, as implemented in LIBLINEAR [16]. We selected Multinomial Naive Bayes as the best model based on the test result, producing the highest AUC, Kappa and weighted average F-Score:

Build the Corpus

Step 1: Collect Tweets

We used the Tweet Search API to harvest 8 million tweets containing one or more query words (e.g. haere mai “welcome” or pōwhiri “reception”) and used these as a query terms. Using the query terms, we produced the highest AUC, Kappa and weighted average F-Score:

Classification results on the test set

We removed all tweets classified as irrelevant, thereby producing the Processed Corpus. A summary of all three corpora is given below:

Table: Corpora

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Raw tweets</th>
<th>TWEETS (0-20)</th>
<th>TWEETS (21-400)</th>
<th>TWEETS (401-200000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>28,800,000</td>
<td>1,628,000</td>
<td>20,000,000</td>
<td>6,000,000</td>
</tr>
<tr>
<td>Processed</td>
<td>28,800,000</td>
<td>1,628,000</td>
<td>20,000,000</td>
<td>6,000,000</td>
</tr>
<tr>
<td>Labelled</td>
<td>49,477</td>
<td>2,495</td>
<td>5,000</td>
<td>1,866</td>
</tr>
</tbody>
</table>

Conclusion

Our search criteria are detailed below:

(2) Ensure tweets are (mostly) written in English.
(3) Convert all characters to lower-case.
(4) Remove retweets and tweets containing URLs.
(5) Remove tweets in which the query word is used as part of a username or mention (e.g. @Happy_kiwi).
(6) The query words containing “whānau” is a compound word for lengthened vowels, search for both the macron and umlaut variants (e.g. Māori and Whānau).
(7) For short phrasal units, search for both the space and non-space variants (e.g. “kia tō” and “kia to”).
(8) Remove tweets containing fewer than five tokens due to insufficient volume analysis.

The resulting collection of tweets, termed the Original Dataset, was used to create the Raw Corpus.