A Hybrid Architecture for Labelling Bilingual Māori-English Tweets

David Trye, Vithya Yogarajan, Jemma König, Te Taka Keegan, David Bainbridge & Mark Apperley
Research Aim

• To improve automatic **language identification** for Māori-English text
  • Focusing on (noisy) Twitter data

Tauranga kaimoana safe to eat bit.ly/1jfkDBM

8:02 PM · Dec 3, 2013 · twitterfeed

Tauranga kaimoana

safe
to
eat

2x Māori

3x English

Bilingual
Contributions

- System
- MET Corpus
- Evaluation
- Vis Tools
Background

• Te reo Māori is often interspersed with English
  • **Code-switching:** ‘multi-word stretches’ (Poplack, 2018)

  Hari huritau ki a koe! Hope you have a wonderful day e hoa!! 💖🎉

  8:50 AM · Sep 2, 2022 · Twitter for iPhone

  • **Loanwords / borrowings:** (mostly) individual words

  Tēnā koe, e hoa! Ngā mihi ki a koe mo retweet!

  Translate Tweet

  3:32 PM · Aug 22, 2018 · Twitter for iPhone

  #TIKANGA A temporary rāhui has been put in place to protect sacred maunga in Ngāti Whatua from harmful firework activities

  7:25 PM · Oct 29, 2022 · Buffer
Related Work

- Identifying Māori-English code-switching points

Key Research Question:
Can these two approaches be fruitfully combined into a hybrid system?

- Extracting Māori text
  - Rule-based approach (Te Hiku Media)
    - RMT System (Trye et al., 2022)
  - Machine learning-based approach
    - ML System (James et al., 2022a)
The Bigger Picture

• Te reo Māori is fundamental to Māori culture

• Both Māori and New Zealand English are under-represented in speech and language technology
  • There is a critical need for new systems and resources to address this (James et al., 2020, 2022a)

• Existing NLP tools are biased towards (certain varieties of) English (Hovy & Prabhumoye, 2021)
  • These tools often fail to recognise or correctly spell/pronounce Māori words (“Kaitaia” → “Car Tyre”)

• Our goal is to reduce this inequity in NLP resources
Key Challenges

• Lexical overlap
  • Both Māori and English use the Roman script
  • 100+ **interlingual homographs**
    • Words that are spelt the same but have different meanings across languages (Dijkstra, 2007)
    • *i, a, hope, here, more, kite*, etc.

• Social media language
  • Internet slang, abbreviations, acronyms
    • *haha, ktk (Māori equivalent of lol), amirite, cuzzie*
  • Misspellings, typos
  • Neologisms
  • Emojis, hashtags, GIFs, etc.
RMT System

• Based on rules by Te Hiku Media
• Tokens must contain valid Māori characters
  • 5 vowels (i, e, a, o, u)
  • 10 consonants (p, t, k, m, n, ng, wh, r, w, h)
• Tokens must follow Māori syllable structure
  • Consonant/vowel alternation: (C)V(V), (C)V₁V₁V₂
  • No consonant clusters
  • End with a vowel
• Lengthened vowels may be indicated with a macron (ā) or double vowels (aa)
ML System

• Bidirectional Gated Recurrent Units (Bi-GRU)
  • Attention layer based on Bahdanau mechanism
  • Trained on Hansard dataset (James et al., 2022b)

• Text represented using fastText word embeddings
  • Skip-gram model with 300 dimensions
  • Pre-trained on Māori & Māori-English corpora (James et al., 2022a)

• Model trained to predict M/E/B tweets
  • Networks optimised with Adam (Kingma and Ba, 2015)
  • Softmax activation in output layer
  • Dropout rate of 0.5 and early stopping used
Pre-processing

• Collected tweets comprising roughly **30-80% Māori text** from known Māori-language users
  • Users identified via *Indigenous Tweets* (Scannell, 2022)

• Tweets were subsequently **cleaned**
  • Stripped non-Roman characters (漢字)
  • Standardised user mentions (@user) & links (<link>)
  • Expanded English contractions (isn’t → is not)

• Discarded ~40,000 **irrelevant tweets**
  • Retweets, bots, duplicates, short tweets (<4 tokens)
  • Tweets containing other languages (not exhaustive)

***"Hi! Heiva i Tahiti! Te ineine mai ra! Pā'oti i ni'a, pā'oti i raro, tīfene, tīfene, 'ami, 'ami"***
Token-Level Labels

Pre-processed Tweets → Tweet Tokens

- RMT System
  - Māori
    - Update label
    - Contextual check of neighbouring tokens
    - Update label
  - $M_{RMT}$

- ML System
  - Ambig.
    - $E_{ML}$
  - Both $M$

- FastText
  - $E_{FT}$

- NLTK
  - $E_{NTLK}$

Remaining Tokens

#hashtags, user mentions, links, emoticons and punctuation

All three $E$
MET Corpus Summary

76,416 tweets

- Māori: 1.1%
- English: 10.3%
- Bilingual: 88.6%

781,381 tokens

- Māori: 40.5%
- English: 59.5%

Limitation: Many tweets were filtered out of the corpus to improve accuracy, such as tweets with one or more ‘Unknown’ or ‘Ambiguous’ labels.

2,417 users

- Bilingual: 2347
- English: 1148
- Māori: 283
Manual Annotation

- We manually labelled **850 tweets** for evaluation purposes
  - All three systems (RMT, ML & Hybrid) at both the token and tweet level
  - Strong agreement between annotators
    - Cohen’s $\kappa = 0.816$ for a subsample of 200 tweets

- Recorded information about each mistake
  - False negative (FN) or false positive (FP)?
  - Specific **error type**
    - Interlingual homograph
    - Named entity (person, place, iwi, organisation, event, etc.)
    - Illegal character(s)
    - Misspelling or missing macron(s)
Evaluating our system

- Hybrid system had the fewest token-level errors, followed by RMT system.

![Bar chart showing False Negatives and False Positives for RMT, ML, and Hybrid systems.]

- Examples include:
  - False Negatives (M misclassified as E): kia, e, o, tau
  - False Positives (E misclassified as M): i, a, Waitangi

- Also best in terms of interlingual homographs.
### Evaluation Metrics

#### Token-Level

<table>
<thead>
<tr>
<th></th>
<th>F1-Score</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>E</td>
<td>M</td>
<td>E</td>
</tr>
<tr>
<td>RMT</td>
<td>0.90</td>
<td>0.87</td>
<td>0.93</td>
</tr>
<tr>
<td>ML</td>
<td>0.94</td>
<td>0.85</td>
<td>0.94</td>
</tr>
<tr>
<td>Hybrid</td>
<td>0.95</td>
<td>0.94</td>
<td>0.94</td>
</tr>
</tbody>
</table>

#### Tweet-Level

<table>
<thead>
<tr>
<th></th>
<th>F1-Score</th>
<th>Specificity</th>
<th>Overall Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>E</td>
<td>M</td>
<td>B</td>
</tr>
<tr>
<td>RMT</td>
<td>0.06</td>
<td>0.39</td>
<td>0.91</td>
</tr>
<tr>
<td>ML</td>
<td>0.71</td>
<td>0.40</td>
<td>0.93</td>
</tr>
<tr>
<td>Hybrid</td>
<td>0.89</td>
<td>0.51</td>
<td>0.95</td>
</tr>
</tbody>
</table>
Wrapping Up

• We devised a **novel system** for labelling Māori/English text
• We used this system to create an **annotated corpus** of 76,000 tweets
• These developments can facilitate further NLP research for Māori and New Zealand English
• This work could also be impactful for research in other low-resourced languages
A Hybrid Architecture for Labelling Bilingual Māori-English Tweets

Check out our interactive visualisation tools:
• https://bilingual-met.github.io/hybrid/
• https://bilingual-met.github.io/hybrid/sample

Contact me
David Trye
dgt12@students.waikato.ac.nz
References


