MĀ TE KUPU TE WHENUA E ORA AI : THE CHALLENGES OF GEOSPATIAL NATURAL LANGUAGE PROCESSING WITH NEW ZEALAND MĀORI

Niloofar Aflaki[1], Kierin Mackenzie[1], Hone Morris[2], Hans W. Guesgen[3], Jon Procter[4] and Kristin Stock[1]

[1]Massey Geoinformatics Collaboratory, Massey University, Albany, New Zealand

[2]Department of Māori Studies, Massey University, Palmerston North, New Zealand, hapu: Ngāi Te Rangitotohu, Ngāti Mārau, Ngāti Maru

[3]School of Mathematical and Computational Sciences, Massey University, Palmerston North, New Zealand

[4]School of Agriculture and Environment, Massey University, Palmerston North, New Zealand, hapu: Muaūpoko

Abstract. Recent years have seen concerted efforts to revitalise New Zealand Māori, the indigenous language of Aotearoa New Zealand, after earlier attempts at suppression during colonisation. Automated methods for natural language processing together with the increasing availability of written Māori language resources have great potential for extracting knowledge from text to increase understanding of current and historical Māori worldviews. In the geographic domain, these methods can be used to increase knowledge of Māori conceptualisations of landscape and to enable information retrieval for purposes such as mapping of species distribution and disaster events. However, most existing tools are based on the form and syntax of English and other well-resourced languages and pose challenges when applied to Māori, including lack of annotated data, inappropriate grammatical assumptions and high levels of polysemy. We discuss these challenges as discovered during (1) the creation of a large Māori corpus through the amalgamation of multiple other language resources; and (2) the comparison of five rule-based and machine learning bag of words methods to identify geographic senses of a collection of 11 geographic feature type words, many of which have multiple other meanings.

Introduction

The field of natural language processing (NLP) has advanced rapidly in recent decades, with the development and refinement of automated methods for the extraction of knowledge from text. These methods provide a basis for **geospatial** natural language processing, which automatically extracts geospatial information from text sources such as reports, documents, blogs and social media. This information may include descriptions of locations and objects or activities that may be found there (e.g., *we collected kai moana near the estuary, opposite the island*) as well as references to specific places and the stories behind them (e.g., *Te Tangihanga ō Kupe* can be translated as the mourning of Kupe in reference to the sound of the waters around it, and *Ngā Rā o Kupe* means Kupe's sails, referring to a story in which Kupe and his companion Ngake competed to make sails, as well as to the shape of the cliffs in that place).

The automated extraction of knowledge from text relies on a range of tools for text processing and analysis, and while these tools are well developed for English and some other widely spoken and well-resourced languages, Māori suffers from a lack of the same level of NLP support. Examples of these tools include part-of-speech tagging, which identifies grammatical elements (nouns, verbs, etc.) in text; relation extraction, which extracts dependencies between words (e.g. the subject and object of a verb); lemmatization, which identifies the root form of a word; and word sense disambiguation, which identifies different senses of words that have

[1] Translates into English as: 'through Māori terminology the land will live.'

multiple meanings (e.g. bank referring to either a river bank or a financial institution). These tools may be used directly to identify and extract specific items of content from text as well as to pre-process text for other forms of analysis such as topic modelling, which enables new insights to be discovered in text.

While work has begun on the development of some NLP tools that are specific to the Māori language, particularly addressing speech recognition (James et al., 2020) and pronunciation (Watson et al., 2017), the application of existing tools for many standard functions poses some challenges for Māori. First, many tools rely on repositories of annotated text for training of their models, which work by learning from examples of text that have been manually classified or tagged. Few annotated resources currently exist for Māori, and efforts to create such resources are more challenging than for many languages due to the relatively small number of fluent Māori language speakers. Second, the most recent and best performing deep learning methods rely on embeddings, reduced dimension vector representations of the semantics of individual words based on the words that are commonly located in their vicinity (Clark et al., 2012). Embeddings provide a representation of the meaning of words that can be automatically extracted and provide a foundation for many other NLP tools. While sets of pre-trained and publicly accessible embeddings created from multiple models exist for English and other well-resourced languages, no such resources are available for Māori. Very large text corpora are necessary to create high quality embeddings, and embeddings can easily be distorted by imbalanced or biased text resources. Māori language corpora of this magnitude are not yet available. A third group of challenges arises due to the inherent differences in the form and structure of the Māori language compared to English and other western languages on which the creation of NLP tools is based. Part-of-speech taggers and lemmatizers assume a particular grammatical structure, and characteristics of Māori such as high levels of polysemy and concatenation of words to produce different meanings mean that common NLP pre-processing methods such as the use of stop-word lists as less effective than for other languages.

The contribution of this paper is to identify and demonstrate these challenges through two activities. First, we describe the creation of a Māori language corpus through the amalgamation of several other corpora containing both current and historical text of multiple types, including historical legal documents, newspapers, web pages and social media. We explain the characteristics of these sources and their impact on efforts to perform large scale natural language processing. Second, we describe an experiment in the use of rule-based and machine learning methods to distinguish geographic senses of 11 words that describe geographic features in the landscape from multiple other senses. Given challenges with manually annotated data that are needed for machine learning approaches, we evaluate the use of simple rule-based approaches which do not require annotation. We show that rule-based methods provide better precision than basic machine learning bag of words methods for geographic feature type terms that refer to specific types of landscape features such as *motu* (island), *moana* (lake, ocean, sea), *puke* (hill, mound) and *puna* (spring).

The paper is structured as follows. Section 2 presents related work in corpus building with low-resourced languages, Māori NLP tools and word sense disambiguation technologies. Section 3 explains the creation of the corpus, including sources, methods and challenges. Section 4 describes the geographic word sense disambiguation experiment, results and challenges. Section 5 provides a conclusion and discusses future work.

Related Work

Corpus Building for Low-Resource Languages

Although some low-resource languages like Urdu are widely spoken (Saeed et al., 2019), others are on the decline, and having quality natural language processing tools can help with revitalisation efforts. Issues can be broadly similar for threatened languages or endangered languages (Neubig et al., 2020). Tools for more well supported languages do not always work for under-resourced languages.

Building a corpus in a low resource language often begins with language identification. Not all material in a language is labelled as such, and there needs to be ways of isolating examples of the target language from other

potentially similar languages (King, 2015). Failure to do so can result in a corpus with unintended language examples.

Web crawling with a language identification module is one way to put together corpora for low-resource languages (Scannell, 2007). Web corpora like these have been used for languages like Mi'kmaq (Maheshwari et al., 2018) One source for corpora for pre-made low resourced languages is Sketch Engine. Their TenTen corpora are crawled from the web using a web spider. There are over 40 languages represented, and many of these are low-resource languages[2].

There has been a concerted effort for creating solid corpora for some low resource languages. The National Corpus of Irish is currently being created by Gaois at Dublin City University. The corpus will have a final projected size of 155 million words, using written and spoken corpora as well as a monitor corpus adding around a million words a year [3]. There is also a National Corpus of Contemporary Welsh, containing over 14 million tokens, and just over 11 million words (Knight et al., 2020) gathered from a wide range of current Welsh in use.

NLP for the Māori language

There has been a sustained effort over the past two decades to build up various tools for NLP of Te Reo. One of the earlier works involved the creation of the Te Kaitito system, which was a collection of NLP resources made of modular components for sentence translation, mixed-initiative dialogue, computer-assisted language learning and dynamic hypertext (Knott et al., 2002). Dialogue was at a basic level but could be interacted with in either Te Reo or English. The dialogue system was eventually integrated with a talking head interface in a conversational agent called Kare (King et al., 2003).

Analysis of contemporary Māori language use was facilitated by the creation of the Māori Broadcast Corpus (Boyce, 2006). The corpus was made of just over one million tokens, and Boyce conducted analysis on generating frequency and distributional information as well as different word senses. A similar approach was taken with the Legal Māori Corpus as part of the Legal Māori Project for the creation of the Legal Māori Dictionary, particularly finding different senses of key terms like Mana (Boyce, 2011).

The MAONZE project is a key oral corpus and looks at changes in the pronunciation of Te Reo Māori over time. The MAONZE teamhas prioritised historical depth over raw size (King et al., 2011). The corpus has been useful for other work in Te Reo, and was a key resource in developing a pronunciation aid for the Māori language (Watson et al., 2017), as well as evolution and trajectory of diphthongs (Stoakes et al., 2019).

Cocks (2012) worked on part-of-speech tagging for Māori as well as diacritic restoration tools for macron restoration. The Māori Macron Restoration service is still online[4]. However, conventional part-of-speech systems do not necessarily work well for Māori, and Māori does not require a noun/verb distinction for individual words. Te Reo can be seen as a phrase based language, and there are two separate syntactic categories for phrases that can be used instead (Yamada, 2014). Other work has been done with machine translation and word alignment using a parallel Māori - English corpus (Mohaghegh and Sarrafzadeh, 2016).

Related language work may point to other approaches. Recent NLP work with Cook Islands Māori has focussed on untrained forced speech alignment, speech-to text models, and part-of-speech tagging (Coto-Solano et al., 2018). Hawaiian, another related language, has a similar problem facing Te Reo where the orthography has changed over time. There has been some work using weighted finite state transducers and a recurrent neural network language model (Shillingford and Parker Jones, 2018).

More recently the Reo Māori Twitter was created from Māori language tweets. Trye et al. (2022) focussed on known users of te reo for their collection as the Twitter API did not have official support for Te Reo. English

[2] https://www.sketchengine.eu/documentation/tenten-corpora/
[3] https://www.corpas.ie/en
[4] http://community.nzdl.org/macron-restoration/jsp/en/main.jsp

tweets and tweets containing less than 70 – 80% Māori were removed, as were formulaic tweets. This appears to be the first social media corpus in Te Reo Māori (Trye et al., 2022).

Te Hiku Media is working on a Te Reo Māori Part-of-speech tagger and has already developed an automatic Māori speech recogniser and development tool (Trye et al., 2022). Te Hiku Media also notes the importance of data sovereignty over Māori language resources. They aim to further develop a range of natural language processing tools for Te Reo Māori as well as other languages used in Aotearoa New Zealand[5]. A Te Reo Māori Text To Speech synthesis system is also being developed, and a Māori speech corpus was developed to facilitate the work (James et al., 2020).

Word Sense Disambiguation

Word sense disambiguation (WSD) is the process of determining the meaning of a word based on its context, usually by means of a computer program. Many words in human language have more than one meaning (e.g., the word bank may refer to the bank of a river, a financial bank, or as a verb, to tilt steeply), but humans find it easy in most cases to determine what the meaning of a word is in a given context. This is not the case for a computer program, which must analyse unstructured textual information to determine the underlying meaning.

Most approaches to WSD come from the field of machine learning. They range from supervised learning approaches to pattern recognition ones:

- **Supervised WSD** (e.g. (Lai et al., 2021))
  In these approaches, labelled training sets are used to train a classifier. The training sets use examples that have been encoded using a fixed set of features. Each example is annotated by a ground truth, which provides the word sense.
- **Unsupervised WSD** (e.g. (Ustalov et al., 2018))
  In these methods, unlabelled corpora are used without any manually added senses.

Generally, supervised approaches to WSD produce the best results. They include approaches based on decision lists, decision trees, naïve Bayes, neural networks, exemplar learning, support vector machines, and ensemble methods. Among the unsupervised approaches are context clustering, word clustering, and co-occurrence graphs. A detailed description of these approaches, in addition to other approaches, is provided in (McCarthy, 2009).

In addition to using a purely supervised or unsupervised approach, some authors opt for a semi-supervised approach (Torunoğlu-Selamet et al., 2020), or they enhance WSD by using additional knowledge about the language (Rouhizadeh et al., 2020). Particularly promising are also deep learning approaches (Saeed et al., 2021) , which combine advantages of supervised and unsupervised learning. They are a feature-based learning approach, but unlike other neural network approaches, they learn features directly from the input.

In a geographic context, word sense disambiguation has been applied to geographic text to detect the sense of spatial relation prepositions. Prepositions such as *near, at* or *beside* describe the spatial relationship between two objects (e.g. *the house beside the river*), and may be used to describe geographic location being the location between objects on the surface of the earth, as well as more generic spatial relationships between objects (e.g. *the cup beside the book*), or in non-spatial senses (e.g. *I was beside myself with joy*). Machine learning methods have been very successful at distinguishing geographic from other spatial senses, and from non-spatial senses for the English language (Radke et al., 2019). Machine learning classifiers have also been applied to distinguish a range of geographic feature type senses in the Māori language (Stock et al., 2019). Geographic feature types are geographic objects in the landscape (e.g., rivers, hills, mountains, lakes etc.). This work uses a bag of words classifier to achieve high precision and recall on a small collection of historical Māori language newspapers

[5] https://www.mbie.govt.nz/science-and-technology/science-and-innovation/funding-information-and-opportunities/investment-funds/strategic-science-investment-fund/ssif-funded-programmes/te-hiku-media/

for ten feature types but relies on a substantial amount of manual annotation. Our work goes beyond this in proposing and evaluating the use of rule-based methods to reduce the requirement for manual annotation.

Corpus Creation

Data sources

We created a large corpus by combining several different existing corpora with a range of data sources and types. These are each described in the following paragraphs, and **Table *1*** presents the number of tokens (total number of words) and unique tokens in each of these corpora and in total.

Nga Tautohetohe Reo | The Hansard Reo Māori Corpus was put together by Te Hiku Media[6] and is a repository of Te Reo Māori utterances in the New Zealand parliament. Contributors have been Caleb Moses, Edward Abraham, William Ti'iti'i Asiata and Tyla Hill Moana. Utterances have a threshold of more than 50% of total words in Te Reo Māori, excluding words that are ambiguous.

The Pre 1910 Legal Māori Corpus comes in several formats. There is an "English Removed" set, an "English Retained" set, and an "English Removed by Text Category" set. We used the last of these. The set was compiled as part of the Legal Māori Project at the Victoria University of Wellington by a team headed by Māmari Stephens and Mary Boyce [7,8]. There is a more complete Legal Māori Corpus that spans from 1829 to 2009, but this can currently only be searched through at the Legal Māori Research Hub[9]. Full texts up to 1910 are also available at the Legal Māori Archive hosted by NZETC.

The Māori Niupepa Collection (Niupepa: Māori Newspapers) is a collection of newspapers published between 1842 and 1932 and is hosted by the New Zealand Digital Library Project at the University of Waikato Department of Computer Science. It covers 17,000 pages and 34 different periodicals. According to statistics given on the website[10], 70% of the writings are in Te Reo Māori, 27% are bilingual, and 3% are in English. It is the largest of the corpora that make up our corpus, and as such occasionally has a disproportional effect.

The Reo Māori Twitter Corpus has been discussed previously, but covers data from 2007-2020 with a peak in 2014, and is "the largest publicly-available collection of social media data containing (almost) exclusively Māori text" (Trye et al., 2022). Overall, 79,000 Māori language tweets are included, and the corpus has been processed to exclude non-Māori tweets and other potential noise.

The Mitenten20 Corpus was obtained from Sketch Engine and was put together from the Maori WaC corpus 2013, crawling of the Māori Web, and Māori Wikipedia. Unlike some of the other Ten Ten corpora, the miTenTen corpus was not part-of-speech tagged or lemmatized[11]. The corpus contains some texts in Japanese and in various languages related to Te Reo Māori, like Hawaiian.

Journal articles and theses were found by searching Google Scholar and New Zealand university repositories for common pairs of Māori words like "I te" "ki te" "o te" and "i nga, "o nga", as well as combing through journals known to have a high concentration of articles in Māori. Closely related languages that share words with Māori were excluded using features that are present in those languages but not Māori, like the letters b and v. Access to articles of He Pukenga Kōrero were organised through Margaret Forster. Other key journals include the Journal of the Polynesian Society, the MAI Journal, the MAI Review, Te Kaharoa, and Te Kōtihitihi. PDFs were downloaded and txt files extracted from the pdfs using PDF-Tools from Tracker Software. Some theses could not be extracted due to being locked and were not used. Some documents required OCR before text extraction.

---

[6] https://github.com/TeHikuMedia/nga-tautohetohe-reo
[7] http://nzetc.victoria.ac.nz/tm/scholarly/tei-legalMaoriCorpus.html
[8] https://www.legalmaori.net/about
[9] https://www.legalmaori.net/corpus
[10] http://www.nzdl.org/cgi-bin/library.cgi?a=p&p=about&c=niupepa
[11] https://www.sketchengine.eu/mitenten-maori-corpus/

Table 1. Total size of the corpus, number of tokens and number of types

Sources	Number of tokens	Number of unique tokens
Hansard Reo Māori corpus	484,932	12,798
Pre 1910 Legal Māori Corpus	4,418,231	43,335
Niupepa: Māori newspapers	11,223,350	135,366
The Reo Māori Twitter (RMT) Corpus	982,370	33,711
Mitenten20 Corpus	7,056,152	95,584
Journal articles	1,310,124	49,522
Theses	2,723,402	64,980
The whole corpus	28,198,563	272,479

Data Preparation

As mentioned earlier, we collected data from several sources, with each source contained different metadata. For example, the Niupepa metadata included publication date, article number, article title, volume, newspaper name and URL. In contrast, the Hansard metadata excluded newspaper name, article title and volume, but included creator. We used the MongoDB NoSQL database management system (Győrödi et al., 2015) as it can manage and support querying of unstructured data with different metadata and multiple languages.

In the process of importing the text and its metadata into MongoDB, we removed all punctuation and non-Latin characters and converted all text lower case. Finally, we removed all macrons. The Māori language uses macrons over some vowels (ā ē ī ō ū) to indicate different pronunciation, and the meanings of words can be changed by the presence of a macron. However, while some of the sources included macrons in their data (some parts of the RMT, Mitenten20 and Hansard corpora and some of the Māori theses), most did not. This inconsistency causes problems for querying and analysis, as it does not recognise words that do not have macrons as the same word as those without (e.g. Māori and Maori). However, this does result in some loss of the ability to distinguish between meanings of words for which macrons validly indicate different meanings (for example *keke* meaning cake vs. *kēkē* meaning armpit).

Discussion

The cleanliness of the source corpora that were used to create our corpus varied. Some had been cleaned entirely of other languages, like the Pre-1910 Legal Māori Corpus. Others had a high degree of other languages, particularly the Mitenten20 Corpus from Sketch Engine. Due to similarities between written Māori, Japanese, and other Pacific languages, like Hawaiian, many Japanese and Hawaiian texts made it into the corpus. Māori and other Polynesian languages alternate consonants with vowels, as does Japanese. Disambiguation of Māori from these other languages can be done to some extent using letters not used in Te Rēo, like b, v, or the ‘okina used in Hawaiian, or words ending in an n, as in Japanese. In future work, these documents will be further filtered. English was also found in several of the corpora used, as many of the texts were bilingual to some degree.

Another complication is that Te Reo Māori has several orthographies. The current standard uses macrons to distinguish between short and long vowels, but this did not become standard until the 1960’s. Older writings use a mix of macrons or other diacritical marks, doubling of vowels, the diaeresis/umlaut or not marking the difference between short and long vowels. There are also dialectical differences, particularly between the South Island and elsewhere (e.g. the term for mountain is *mauka* on the South Island and *maunga* elsewhere).

Working with low-resource languages has some inherent difficulties compared to working with well-resourced languages like English or Spanish. These include a lack of annotated data, a high degree of bilingualism in sources like tweets (Agüero Torales, 2022), and fewer available pre-existing lexical resources (Lind et al., 2019). Available text materials are reduced, making generalisations from the corpus to the language as a whole difficult, as the corpus is less likely to be a representative or balanced sample (Vinogradov, 2016). Very few

languages have speech regulation and machine translation available, machine-readable dictionaries, thesauri etc (Scannell, 2007).

In addition, tools and approaches that are commonly used for well-resourced languages may not be suitable for some low-resource languages like Māori. For instance, stop word lists are commonly used in English to exclude words that are grammatical but do not have lexical meanings (e.g. the, and), but in Māori, many of the words that might usually be placed on a stop words list also had lexical meanings. This means that a simple list might not be the best approach to filtering out non lexical words, at least not without some part-of-speech work to disambiguate non-lexical from lexical uses. One approach that would still allow for a stop words list would be to only include words that are rarely used for lexical purposes. Other approaches may be more accurate.

There can be advantages with the grammar of some low-resource languages as well. Lemmatizers are used in English to obtain the root form of words after removing plural markers and verb conjugations, etc. The need for lemmatisation in Te Reo Māori is less important as there are fewer different word forms, and these tend to be limited to verbs in their passive form (hāngūtanga), which take on the following endings: *-tia, -hia, -ngia, -a, -ia, -ina, -kia, -mia, -na, -nga, -ria, -whia, -whina, -kina*[12].

Geographic Feature Type Sense Disambiguation

Having identified several challenges involved in using NLP with the Māori language that became apparent in the process of creating the corpus, we now conduct an experiment to demonstrate the application of NLP to Māori text, and to evaluate the use of rule-based methods to detect the geographic sense of geographic feature type terms. While machine learning methods have been shown to generally perform better than rule-based methods for NLP, the lack of annotated resources for Māori means that rule-based methods may be more practical for some purposes.

Geographic feature types are geographic objects in the landscape (e.g. rivers, hills, lakes), and we refer to words for these geographic feature types as geographic feature type words. However, due to polysemy, geographic feature type words may also have other senses, and the purpose of this experiment is to distinguish the geographic from non-geographic senses. This is particularly challenging for Māori, due to high levels of polysemy, and for the common metaphoric use of geographic feature type words. For example, the word *motu* refers to an island as an area of land in the ocean or a lake, but also to anything that is isolated or separated, including a clump of trees, and the same word also has many verb senses meaning to sever, separate, cut off, free or escape[12]. Determining the uses of the word *motu* that are nouns and that refer to a physical geographic object rather than other senses can be challenging, and while the part of speech can be useful evidence to assist in this task, part-of-speech taggers for Māori are still not readily accessible, particularly with the range of options and models that are offered for English and other well-resourced languages (Toutanova et al., 2003). The models and the training sets used to create NLP tools influence the accuracy of their results, and work is needed to develop similar tools for Māori. Furthermore, the tag sets used for English and other well-resourced languages are not appropriate for Māori, with some work already conducted to develop more suitable tag sets to incorporate into part-of-speech taggers (Cocks, 2012).

Methods

In this paper, we used four methods to disambiguate geographic feature types (i.e. distinguish geographic senses from non-geographic senses of geographic feature type words). We evaluated the methods using data from two corpora:

Corpus 1: A corpus described in (Stock et al., 2019), consisted of a collection of early issues of two newspapers in the Māori language: Te Puke Ki Hikurangi[13] and Te Ao Hou[14] and included the first 20 issues of Te Ao Hou, covering the period from 1952 to 1957 inclusive and the first 10 issues from Te Puke Ki

[12] https://maoridictionary.co.nz/
[13] http://www.nzdl.org/cgi-bin/library?a=p&p=about&c=niupepa
[14] https://paperspast.natlib.govt.nz/periodicals/te-ao-hou

Hikurangi, all from 1897. These newspapers were selected because they were written and managed by Māori. This corpus in total consisted of 794,649 word tokens and 29,837 word types.

Corpus 2: The corpus described in Section 3.

We extracted all instances of five geographic feature type terms from Corpus 1 and the remaining six terms from Corpus 2. The geographic feature type terms were selected from the list provided in (New Zealand Geographic Board, Ngā Pou Taunaha O Aotearoa, 2014), which contains 354 terms. We selected the most frequently occurring 19 geographic feature type words in Corpus 2, excluding those that had very few geographic senses on manual examination by one of the co-authors who is a frequent Māori speaker (Morris). For our experiments, we created a data set containing 200 annotated instances of each of the these most frequent 19 terms. Seven of the geographic feature type terms had previously been annotated as either a geographic sense of the term (class 1) or a non-geographic sense (class 0) in Corpus 1 as described in (Stock et al., 2019), and we randomly selected 200 instances from each of these terms. All instances of the remaining 13 geographic feature terms were extracted from Corpus 2 and 200 instances were randomly selected. The same co-author (Morris) then annotated each term in the same way as for Corpus 1, in discussion with one other co-author (Mackenzie) to resolve ambiguous cases. When information was inadequate to determine whether the term was being used in a geographic or non-geographic sense, the case was placed in class 0. This occurred because the window of words did not contain sufficient words to detect the context and thus determine whether the word was geographic, it was illegible or in a different language.

Having completed annotation of all 20 terms, we then excluded all terms that had heavily imbalanced splits of instances class 0 and class 1, meaning that fewer than 10% belonged to either of these classes, as such skewed data sets distort the results from method evaluation. The remaining 11 terms that were used in our experiments are shown in **Table *2*** For each term, we extracted windows containing 10 words on either side of the geographic feature type term and use them in the following methods.

Table 2. Geographic feature type words used in word sense disambiguation experiment

Māori	**English translation**	**Corpus**
motu	island	1
whenua	land	1
wāhi	place	1
kāinga	home, village	1
puna	spring	1
wai	water	2
moana	lake, sea, ocean	2
tai	tide, ocean, sea, coast	2
pae	Region	2
puke	hill, mound	2
uta	inland, shore, interior	2

Method 0: All Geographic Senses

Method 0 is provided as a baseline for comparison of the other methods and assumes that every geographic feature type term is a geographic sense (class 1). This means that every actual geographic sense will be identified by this method, while many senses that are not geographic will be incorrectly identified as geographic. The success of this method depends only on the proportion of geographic senses of the word (the % of the data set for a given geographic feature type term that are classed as 1).

Method 1: Collocated Geographic Feature Type Frequency

Method 1 counts the frequency of geographic feature type terms in a 10-word window (on either side of the word of interest). We use the full list of 354 Māori geographic feature type terms found in (New Zealand Geographic Board, Ngā Pou Taunaha O Aotearoa, 2014). We consider a geographic feature type term to be geographic if it is over a specified threshold. We test a range of thresholds from 1 to 5 (5 meaning there are 5 geographic terms in the 10 word window).

Method 2: Collocated Definitive Particle + Geographic Feature Type Frequency

The second method is same as Method 1 except that we only count a geographic feature type term in the 10-word window if is immediately preceded by a definitive particle (indicating that it is most likely a noun). We used 25 definitive particle terms in Māori (*te, ngā, tētahi, ētahi, tēhea, ēhea, tēnei, ēnei, tēnā, ēnā, tērā, ērā, taua, aua, he, nga, tetahi, etahi, tehea, ehea, tenei, enei, ena, tera, era*). As for Method 1, a given instance of a geographic feature type term is considered geographic if the number of definitive particle+geographic feature type term pairs within the 10 word window is over a specific threshold, and again we test between 1 and 5.

Method 3: Weighted Collocated Geographic Feature Type

In Method 3 we weight the geographic feature type terms within the window using two measures:

1. Distance from the word of interest, on the basis that other geographic feature type terms that are close to the word of interest are more likely to indicate that it is geographic than those that are further away and
2. Likelihood that a geographic feature type term is geographic, determined manually as described below, with the goal of giving greater influence to those terms that almost certainly use a geographic sense, than those that almost certainly do not.

We calculate a geographic index for each instance using the equation 1.

$$geographic\ index^{s} = \sum_{k=0}^{n} \frac{v^{s} * (\frac{1}{7})}{number\ of\ words\ between\ s\ and\ g+1} \qquad (1)$$

Where (s) is the geographic feature type word that we want to determine the sense of, (g) is some other geographic feature type word in the window 10 words on either side of the subject word (s) and (v) is a manually assigned value for the gft in range 1-7 indicating likelihood that the word is geographic (see below). So, if the geographic feature type is closer to the middle term, the weight is higher than if it is distant. Again, if the score for a given geographic feature type term is over a specified threshold, we consider it a geographic sense. We tested thresholds between 0 and 1, incremented by 0.02.

The value (v) was assigned for each of the most frequently appearing 271 terms on the list of 354 (New Zealand Geographic Board, Ngā Pou Taunaha O Aotearoa, 2014). We did not consider those that appeared fewer than five times in the corpus. Assignment of the value for (v) was based on an examination of texts that use the word to indicate its range of geographic and non-geographic senses. We used a seven term Likert-scale and a fluent Māori speaking co-author (Morris) assigned a score based on the question "is this used as a geographical term?" with the following options:

1: Never

2: Almost never

3: Occasionally

4: Frequently

5: Usually

6: Almost Always

7: Always

Method 4: Bag of Words

Method 4 used a simple bag of words (Zhang et al., 2010) model with the 10 word-window around the geographic feature type term as input to generate the bag of words. We used term frequency-inverse document frequency (tf-idf) (Salton and Buckley, 1987) values in the bag of words model with the first 1000 most frequent words to classify the samples using two classification methods: SVM and Naive Bayes with the Weka[15] tool.

Results

We evaluate the five methods described in Section 4.1 with 200 instances of each of the 11 geographic feature type terms shown in **Table *2***. **Table 3** shows the results for our three rule-based methods (Methods 1-3) and two bag of words methods (using Naïve Bayes and SVM models), as well as Method 0, and also shows the counts of instances of class 0 (non-geographic sense) and class 1 (geographic sense), and the percentage of all instances that are geographic for each word, as this influences the results. The figures for Methods 1 to 3 are those for the threshold that gave the highest f1 value, and the relevant threshold for each geographic feature type term is shown in the left-most column in the section for each Method (headed 't').

In **Table 3**, the best performing methods for precision (pink), recall (green) and f-measure (blue) are marked, but we exclude method 0 from the selection of best performance, since its performance is dependent exclusively on the percentage of geographic senses, and it always has a recall of 1 because every instance is considered a geographic sense.

As can be seen, Method 4 (bag of words), with either the naïve bayes or SVM models, performs better than the rule-based methods in determining f-measure in nearly all cases. However, the rule-based methods perform better than the bag of words methods for precision in all but four cases. Method 1 and 3 give very similar results, despite the fact that Method 3 includes the refinements of weighting of geographic feature type words that appear in the window by both distance and likelihood that the term is geographic. We anticipate that further fine tuning of the weighting model could result in improvements to Method 3 over the more basic Method 1, which simply counts geographic feature type word frequency. Method 2 does not perform as well as Methods 1 and 3. Method 2 refines the count of geographic feature type terms in the window surrounding the term of interest by excluding those that are not immediately preceded by a definitive particle. This is intended to reduce the likelihood of words that are not nouns being counted on the basis that they are most likely not geographic feature types, but other senses of the geographic feature type words. Given the absence of part-of-speech tagging and other related NLP tools for the Māori language, more sophisticated methods for collecting evidence of the likelihood of a geographic sense were not possible, and the definitive particle was used as an approximation.

The rule-based methods perform best at low thresholds, with the best f1 being achieved for threshold 1 for all of the words for Method 2, and all but one for Method 1, and for the lowest threshold of 0.02 for all but two of the words for Method 3. This means that for the rule-based methods, the best results are achieved when even geographic feature type words with only one other geographic feature type word in the window are considered a geographic sense, and that further restriction by requiring more geographic feature type words in the window deteriorates performance in most cases (*puke* is an exception to this, performing best with higher thresholds for Methods 1 and 3).There is a high correlation between f-measure and the frequency of geographic senses of the geographic feature type word (this is, the more geographic senses of the word there are, the better the performance of the model), particularly for Method 4 (Pearson correlation coefficient 0.87 for Method 4 with SVM), but much lower for the rule-based methods. This highlights the reliance of the bag of words methods particularly on training data and makes Method 4 less suitable for geographic feature type words that have a very low frequency of geographic senses.

[15] https://www.cs.waikato.ac.nz/ml/weka/

Considering specific geographic feature types, the best results overall are achieved for *motu* (island), *whenua* (land), *kāinga* (home, village) and *moana* (lake, ocean, sea), all of which have high frequency of geographic senses. For the geographic feature type word *pae* (region), the rule-based methods perform much worse relative to the bag of words method than might be expected for both precision and recall (although all methods perform relatively poorly due to low proportion of class 1 instances). Pae has a wide range of non-geographic uses when combined with other words (e.g., *pae tukutuku* can refer to web sites), and the bag of words method is able to model these negative examples, while our rule-based methods are not.

In contrast, the rule-based methods perform much better than the bag of words method for *puke* (hill, mound). The results for *puke* are distorted by its common appearance in the Niupepa corpus as part of the title of a particular newspaper (Te Puke Ki Hikurangi). Furthermore, ***Fig. 1*** shows that the rule-based methods provide better precision for geographic feature type terms that refer to specific types of geographic objects, including *motu* (island), *moana* (lake, ocean, sea), *puke* (hill, mound) and *puna* (spring).

The specific nature of these geographic feature types may mean that they are often collocated in text with mention of other geographic feature types, thus resulting in better performance for the rule-based methods, which explicitly look for these terms, in contrast to the Bag of Words method which create more generalised models of any collocated terms.

Discussion

While the bag of words methods perform best in balancing precision and recall in contrast to the rule-based methods, they rely on training data, and perform best when the training data includes a good balance between instances with different class values. However, training data is particularly difficult to obtain for low-resource languages, and while the number of Māori speakers is increasing due to language revitalisation, the availability of fluent speakers to perform data annotation tasks is limited. The experiment also shows clear differences among geographic feature type words and given that there are hundreds of these types of landscape words in Māori, the requirement for manual annotation to reliably detect their geographic senses for geospatial NLP would be substantial. The results of this experiment show that good precision can be achieved with rule-based methods which require no annotation and could be applied across a range of geographic feature type with minimal additional effort. For geospatial NLP tasks that aim to increase understanding of the use of Māori language and the conceptual models that underly landscape terms, precision is sufficient, as it is not necessary to identify every instance of a word for such studies. Other tasks may require the better balance between precision and recall that is afforded by bag of words methods, and in this case, additional annotation resources are required.

It is likely that the availability of additional Māori NLP resources would increase the accuracy of geographic word sense disambiguation. First, the availability of an accurate, well-trained Māori part-of-speech tagger could improve the results from rule-based methods by enabling non-noun uses of geographic feature type terms to be identified and eliminated from consideration. Second, high quality embeddings created from a large Māori language corpus would enable new methods for word sense disambiguation to be applied, including unsupervised methods that cluster embeddings to identify senses.

The importance of a large, well-balanced Māori language corpus is also important for removing the influence of individual sources (for example, as occurred for the word puke, the models of which were distorted by the common occurrence of the newspaper title *Te Puke Ki Hikurangi* in the corpus). If the corpus is not sufficiently large, it is likely that any analysis regarding Māori world views and culture may be influenced by corpus content that responds to societal and political events.

Fig. 1. Geographic word sense disambiguation - precision for different methods

Table *3*. Geographic word sense disambiguation results

feature type	Number of instances			Method 0 all geographic			Method 1 gft term count				Method 2 dp+gft term count				Method 3 weighted gft term count				Method 4 - NB Bag of Words			Method 4 - SVM Bag of Words		
Māori	**0**	**1**	**%1**	**p**	**r**	**f1**	**t**	**p**	**r**	**f1**	**t**	**p**	**r**	**f1**	**t**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**
motu	63	137	69%	0.685	1.000	0.813	1	0.884	0.445	0.592	1	**0.952**	0.146	0.253	0.02	0.884	0.445	0.592	0.881	0.861	0.871	0.901	**0.885**	**0.893**
whenua	54	146	73%	0.730	1.000	0.844	1	**0.952**	0.274	0.426	1	0.950	0.130	0.229	0.02	**0.952**	0.274	0.426	0.909	0.959	0.933	0.923	**0.979**	**0.950**
wāhi	138	62	31%	0.310	1.000	0.473	1	0.757	0.452	0.566	1	**0.789**	0.242	0.370	0.02	0.757	0.452	0.566	0.703	0.726	0.714	0.754	**0.790**	**0.772**
kāinga	54	144	73%	0.727	1.000	0.842	1	0.754	0.299	0.428	1	0.724	0.146	0.243	0.02	0.754	0.299	0.428	0.858	0.924	0.890	**0.875**	**0.972**	**0.921**
puna	172	28	14%	0.140	1.000	0.246	1	**0.700**	0.250	0.368	1	0.667	0.143	0.235	0.02	**0.700**	0.250	0.368	0.429	**0.536**	**0.476**	0.615	0.286	0.390
wai	138	62	31%	0.310	1.000	0.473	1	0.495	**0.742**	0.594	1	0.549	0.452	0.496	0.02	0.495	**0.742**	0.594	**0.586**	0.661	**0.621**	0.565	0.419	0.481
moana	26	174	87%	0.870	1.000	0.930	1	0.940	0.626	0.752	1	**0.957**	0.385	0.549	0.02	0.938	0.609	0.739	0.901	0.885	0.893	0.872	**0.983**	**0.924**
tai	53	147	74%	0.735	1.000	0.847	1	0.701	0.463	0.557	1	0.787	0.252	0.381	0.02	0.701	0.463	0.557	0.824	0.857	0.840	**0.840**	**0.932**	**0.884**
pae	166	34	17%	0.170	1.000	0.291	1	0.187	0.412	0.257	1	0.361	0.382	0.371	0.06	0.194	0.382	0.257	0.512	**0.618**	**0.560**	**0.778**	0.412	0.538
puke	163	37	19%	0.185	1.000	0.312	2	0.458	0.297	0.361	1	0.246	0.405	0.306	0.16	**0.567**	**0.459**	**0.507**	0.333	0.351	0.342	0.450	0.243	0.316
uta	65	135	68%	0.675	1.000	0.806	1	**0.835**	0.600	0.698	1	0.817	0.363	0.503	0.02	**0.835**	0.600	0.698	0.785	0.785	0.785	0.806	**0.859**	**0.832**

Conclusions

This paper has discussed several challenges that arise when applying NLP tools to the Māori language as a result of both the need for large repositories of annotated data which are currently not available and the substantial differences in the grammar and syntax of Māori relative to other more highly resourced languages for which tools are well developed (especially English). We identified these challenges through the creation of a large Māori language corpus, and through the application of geographic word sense disambiguation. In particular, we demonstrated the use of rule-based methods that do not rely on extensive annotation and that could be applied to multiple geographic feature type terms.

The paper highlights a requirement for additional work to support Māori language NLP, with a particular focus on developing high accuracy NLP tools such as part-of-speech taggers and embeddings, as well as the need for heavily modified or entirely different tools for some aspects of NLP, since common NLP tools used for other languages including lemmatizers and stop word lists are not suitable due to the characteristics of the Māori language.

This additional work will enable new knowledge to be extracted from the large volume of current and historical Māori language resources, increasing understanding of Māori worldviews generally, and also specific aspects of Māori knowledge about landscape, place names and geography.

Acknowledgements

This work was funded through the MBIE Endeavour Research Programme “He Tatāi Whenua: A Te Ao Māori landscape classification”.

References

Agüero Torales, M.M., 2022. Machine Learning approaches for Topic and Sentiment Analysis in multilingual opinions and low-resource languages: From English to Guarani.

Boyce, M., 2011. Mana Aha? Exploring the Use of Mana in the Legal Māori Corpus. VUWLR 42, 221. https://doi.org/10.26686/vuwlr.v42i2.5136

Boyce, M.T., 2006. A corpus of modern spoken Māori. Victoria University of Wellington, Wellington, New Zealand.

Clark, A., Fox, C., Lappin, S., 2012. The handbook of computational linguistics and natural language processing. John Wiley & Sons.

Cocks, J., 2012. Diacritic Restoration and the Development of a Part-of-Speech Tagset for the Māori Language (Thesis). University of Waikato.

Coto-Solano, R., Nicholas, S.A., Wray, S., 2018. Development of Natural Language Processing Tools for Cook Islands Māori, in: Proceedings of the Australasian Language Technology Association Workshop 2018. Presented at the ALTA 2018, Dunedin, New Zealand, pp. 26–33.

Győrödi, C., Győrödi, R., Pecherle, G., & Olah, A. (2015). A comparative study: MongoDB vs. MySQL. 1–6.

James, J., Shields, I., Berriman, R., Keegan, P.J., Watson, C.I., 2020. Developing resources for Te Reo Māori text to speech synthesis system. Presented at the International Conference on Text, Speech, and Dialogue, Springer, pp. 294–302.

King, B.P., 2015. Practical Natural Language Processing for Low-Resource Languages.

King, J., Maclagan, M., Harlow, R., Keegan, P., Watson, C., 2011. The MAONZE project: Changing uses of an indigenous language database.

King, S.A., Knott, A., McCane, B., 2003. Language-driven nonverbal communication in a bilingual conversational agent. Presented at the Proceedings 11th IEEE International Workshop on Program Comprehension, IEEE, pp. 17–22.

Knight, D., Morris, S., Fitzpatrick, T., Rayson, P., Spasić, I., Thomas, E.M., 2020. The National Corpus of Contemporary Welsh: Project Report | Y Corpws Cenedlaethol Cymraeg Cyfoes: Adroddiad y Prosiect.

Knott, A., Bayard, I., De Jager, S., Wright, N., 2002. An architecture for bilingual and bidirectional nlp, in: Proceedings of the 2nd Australasian Natural Language Processing Workshop (ANLP 2002). Citeseer.

Lai, H.-L., Hsu, H.-L., Liu, J.-S., Lin, C.-H., Chen, Y., 2021. Supervised Word Sense Disambiguation on Taiwan Hakka Polysemy with Neural Network Models: A Case Study of BUN, TUNG and LAU 11.

Lind, F., Eberl, J.-M., Galyga, S., Heidenreich, T., Boomgaarden, G., Jiménez, B.H., Berganza, R., 2019. A Bridge Over the Language Gap: Topic Modelling for Text Analyses Across Languages for Country Comparative Research 37.

Maheshwari, A., Bouscarrat, L., Cook, P., 2018. Towards Language Technology for Mi’kmaq, in: LREC 2018, Eleventh International Conference on Language Resources and Evaluation. p. 5.

McCarthy, D., 2009. Word sense disambiguation: An overview. Language and Linguistics compass 3, 537–558.

Mohaghegh, M., Sarrafzadeh, A., 2016. Parallel Text Identification Using Lexical and Corpus Features for the English-Maori Language Pair, in: 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA).

Presented at the 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 910–915. https://doi.org/10.1109/ICMLA.2016.0163

Neubig, G., Rijhwani, S., Palmer, A., MacKenzie, J., Cruz, H., Li, X., Lee, M., Chaudhary, A., Gessler, L., Abney, S., Hayati, S.A., Anastasopoulos, A., Zamaraeva, O., Prud'hommeaux, E., Child, J., Child, S., Knowles, R., Moeller, S., Micher, J., Li, Y., Zink, S., Xia, M., Sharma, R.S., Littell, P., 2020. A Summary of the First Workshop on Language Technology for Language Documentation and Revitalization. arXiv:2004.13203 [cs].

New Zealand Geographic Board, Ngā Pou Taunaha O Aotearoa, 2014. Generic Geographic Features Listing – Maori and English (A1685532 - Version 1: June 2014). New Zealand Geographic Board | Ngā Pou Taunaha O Aotearoa, Wellington, New Zealand.

Radke, M., Das, P., Stock, K., Jones, C.B., 2019. Detecting the Geospatialness of Prepositions from Natural Language Text (Short Paper), in: Timpf, S., Schlieder, C., Kattenbeck, M., Ludwig, B., Stewart, K. (Eds.), 14th International Conference on Spatial Information Theory (COSIT 2019), Leibniz International Proceedings in Informatics (LIPIcs). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, p. 11:1-11:8. https://doi.org/10.4230/LIPIcs.COSIT.2019.11

Rouhizadeh, H., Shamsfard, M., Rouhizadeh, M., 2020. Knowledge Based Word Sense Disambiguation with Distributional Semantic Expansion for the Persian Language, in: 2020 10th International Conference on Computer and Knowledge Engineering (ICCKE). Presented at the 2020 10th International Conference on Computer and Knowledge Engineering (ICCKE), pp. 329–335. https://doi.org/10.1109/ICCKE50421.2020.9303675

Saeed, A., Nawab, R.M.A., Stevenson, M., 2021. Investigating the Feasibility of Deep Learning Methods for Urdu Word Sense Disambiguation. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 21, 38:1-38:16. https://doi.org/10.1145/3477578

Saeed, A., Nawab, R.M.A., Stevenson, M., Rayson, P., 2019. A word sense disambiguation corpus for Urdu. Lang Resources & Evaluation 53, 397–418. https://doi.org/10.1007/s10579-018-9438-7

Salton, G., Buckley, C., 1987. Term weighting approaches in automatic text retrieval. Cornell University.

Scannell, K.P., 2007. The Crьbadcn Project: Corpus building for under-resourced languages. Cahiers du Cental 5, 1.

Shillingford, B., Parker Jones, O., 2018. Recovering Missing Characters in Old Hawaiian Writing, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Presented at the EMNLP 2018, Association for Computational Linguistics, Brussels, Belgium, pp. 4929–4934. https://doi.org/10.18653/v1/D18-1533

Stoakes, H.M., Watson, C.I., Keegan, P.J., Maclagan, M.A., King, J., Harlow, R., 2019. The Dynamics of Closing Dipthong Formant Trajectories in Te Reo Māori, in: Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia. pp. 989–993.

Stock, K., Morris, H., Forster, M., Paraku, R., Egorova, E., 2019. He Tatai Whenua: Automated Extraction of Landscape Terms and their Meanings in New Zealand Maori. Presented at the Geocomputation 2019 - Adventures in GeoComputation, Queenstown, NZ.

Toutanova, K., Klein, D., Manning, C.D., Singer, Y., 2003. Feature-rich part-of-speech tagging with a cyclic dependency network, in: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, NAACL '03. Association for Computational Linguistics, USA, pp. 173–180. https://doi.org/10.3115/1073445.1073478

Torunoğlu-Selamet, D., İnceoğlu, A., & Eryiğit, G. (2020). Preliminary Investigation on Using Semi-Supervised Contextual Word Sense Disambiguation for Data Augmentation. 337–342.

Trye, D., Keegan, T.T., Mato, P., Apperley, M., 2022. Harnessing Indigenous Tweets: The Reo Māori Twitter corpus. Lang Resources & Evaluation. https://doi.org/10.1007/s10579-022-09580-w

Ustalov, D., Teslenko, D., Panchenko, A., Chernoskutov, M., Biemann, C., Ponzetto, S.P., 2018. An Unsupervised Word Sense Disambiguation System for Under-Resourced Languages. arXiv:1804.10686 [cs].

Vinogradov, I., 2016. Linguistic corpora of understudied languages: Do they make sense? RK 40, 127–141. https://doi.org/10.15517/rk.v40i1.24143

Watson, C.I., Keegan, P.J., Maclagan, M.A., Harlow, R., King, J., 2017. The Motivation and Development of MPAi, a Māori Pronunciation Aid, in: Interspeech 2017. Presented at the Interspeech 2017, ISCA, pp. 2063–2067. https://doi.org/10.21437/Interspeech.2017-215

Yamada, F.S., 2014. Māori as a Phrase-Based Language (PhD Thesis). University of Hawai'i at Mānoa.

Zhang, Y., Jin, R., Zhou, Z.-H., 2010. Understanding bag-of-words model: a statistical framework. International Journal of Machine Learning and Cybernetics 1, 43–52.

Keywords

Geospatial language, Natural language processing, word sense disambiguation, Māori