Towards Language Technology for Mi’kmaq

Anant Maheshwari,[1] **Léo Bouscarrat,**[2] **Paul Cook**[3]
1. Computer and Information Science, University of Pennsylvania
2. École des Mines de Saint-Étienne
3. Faculty of Computer Science, University of New Brunswick
anantm95@seas.upenn.edu, leo.bouscarrat@etu.emse.fr, paul.cook@unb.ca

Abstract

Mi’kmaq is a polysynthetic Indigenous language spoken primarily in Eastern Canada, on which no prior computational work has focused. In this paper we first construct and analyze a web corpus of Mi’kmaq. We then evaluate several approaches to language modelling for Mi’kmaq, including character-level models that are particularly well-suited to morphologically-rich languages. Preservation of Indigenous languages is particularly important in the current Canadian context; we argue that natural language processing could aid such efforts.

Keywords: Indigenous languages, web corpora, language modelling

1. The Need for Mi’kmaq Language Technology

For over one hundred years the Indian residential school system in Canada forcibly removed Indigenous children from their families, subjected them to physical and sexual abuse, and attempted to erase their Indigenous identities, in part by prohibiting the use of Indigenous languages (Truth and Reconciliation Canada, 2015). Thousands of children died while in this system. The last of these schools did not close until 1996. The Indian residential school system has been referred to as a form of cultural genocide.

On 11 June 2008 then-Prime Minister of Canada Stephen Harper issued an official apology to the survivors of Indian residential schools.[1] The Truth and Reconciliation Commission of Canada was formed to document events surrounding these schools, and to identify ways to improve conditions for Indigenous peoples. In December 2015 the Truth and Reconciliation Commission of Canada released ninety-four calls to action to facilitate the process of reconciliation (Truth and Reconciliation Canada, 2015). Amongst these is the principle that “Aboriginal languages are a fundamental and valued element of Canadian culture and society, and there is an urgency to preserve them.”

Natural language processing (NLP) could help to play a role in the preservation of Indigenous languages by building language technology tools — such as spelling checkers, next word prediction systems, and machine translation systems — to aid in using Indigenous languages in computer-mediated communication.

Mi’kmaq is an Eastern Algonquian language, spoken primarily in Eastern Canada. It is a polysynthetic, free word-order language (Johnson, 1996). Rand (1888) demonstrates the morphological richness of Mi’kmaq with the word *yăle-oole-mâktāwe-p*c*kŏse*, meaning “I am walking about, carrying a beautiful black umbrella over my head.”. Although it has a rich oral tradition, various Roman scripts have been introduced for Mi’kmaq (Battiste, 1985), differing primarily in their representations of vowel length. Roughly 8,000 people reported Mi’kmaq as their mother tongue in the Canada 2011 Census.[2] Mi’kmaq is the most widely spoken Indigenous language in the province of New Brunswick.

Although there exist Mi’kmaq dictionaries (Rand, 1888; DeBlois, 1996) and translated texts (DeBlois, 1990), Mi’kmaq remains a low-resource language. In particular, no (large) machine-readable corpora are available for Mi’kmaq.

There has been very little work in computational linguistics or NLP on Mi’kmaq. The Crúbadán project (Scannell, 2007)[3] built web corpora for over 400 writing systems (as proxies for languages) of which Mi’kmaq was one.[4] These corpora are not publicly available. Brown (2014) studied language identification for over 1300 languages, with Mi’kmaq being included amongst these.

To the best of our knowledge, this is the first computational work to focus specifically on Mi’kmaq.[5] The long-term goals of this research are to 1.) create a large Mi’kmaq corpus to support corpus linguistic studies of, lexicographical analysis of, and training NLP systems for, Mi’kmaq; and 2.) to build a suite of NLP tools for Mi’kmaq, which could potentially contribute to language preservation. In this preliminary work, in Section 2 we first build a web corpus of Mi’kmaq, using methods similar to Scannell (2007), and analyze it. Language models are an important component of systems for many NLP tasks, including spelling correction and machine translation. In Section 3 we evaluate several approaches to language modelling for Mi’kmaq, including character-level approaches that could capture the morphological-complexity of Mi’kmaq. We conclude in Section 4 by discussing directions for future work.

[1] https://www.aadnc-aandc.gc.ca/eng/1100100015644/1100100015649

[2] http://www12.statcan.gc.ca/census-recensement/2011/as-sa/98-314-x/98-314-x2011003_3-eng.cfm

[3] http://crubadan.org/

[4] The number of writing systems included in this project has since climbed to over 2000.

[5] The Truth and Reconciliation Commission of Canada’s calls to action also state that “The preservation, revitalization, and strengthening of Aboriginal languages and cultures are best managed by Aboriginal people and communities.” None of the authors of this work are Aboriginal people. We have consulted with the Mi’kmaq-Wolastoqey Centre at the University of New Brunswick in carrying out this research.

2. A Mi'kmaq Web Corpus

Baroni and Bernardini (2004) describe an approach to automatically creating topically-focused corpora from the web, whereby tuples are randomly formed from a user-created keyword list for the topic of interest, and these tuples are then sent as queries to a search engine. The top-n results for these queries are downloaded, and then post-processed to remove mark-up and boilerplate content, documents in unwanted languages, and duplicate documents, with the final result being a topically-focused corpus. This approach to corpus construction remains widely-used, and is incorporated into commercial lexicographical tools (Baroni et al., 2006).[6]

The method of Scannell (2007) for creating corpora for specific languages begins with an approach quite similar to that of Baroni and Bernardini (2004), except that the search engine queries consist of a high frequency word in the language of interest combined with other words in that language. The words used in the queries are further controlled to avoid words that happen to also occur in another language (e.g., *die* is both an English and German word). The search engine results include many documents in the intended language.[7]

2.1. Corpus Construction

In this section we describe our approach to creating a Mi'kmaq web corpus, which uses an approach similar to that of Scannell (2007).

Seed Word Selection The Universal Declaration of Human Rights (UDHR) is available in over five-hundred languages, including Mi'kmaq.[8] We selected as seed words those words that occur in the Mi'kmaq translation of the UDHR, available through NLTK (Bird et al., 2009).[9]

Although the full text corpora created by Scannell (2007) are not publicly-available, various summaries of the corpora are provided, including a list of the word types, and their corresponding frequencies, in each corpus. In preliminary experiments we considered basing our seed words on the word frequency list provided for Mi'kmaq; however, we found the resulting corpora to be much smaller than when using seed words derived directly from the UDHR (which Scannell (2007) also considered as a source for seed words). We therefore only consider the seed words from the UDHR in the remainder of this paper.

Query Generation We considered two approaches to forming queries from the list of seed words, which were then sent to a commercial search engine.

Random We used the BootCaT tools (Baroni and Bernardini, 2004)[10] to randomly select 3-tuples from the seed word list, and then used these as queries.

Crúbadán Our second approach to query generation is based on that of Scannell (2007). We divide the seed words into high and low frequency terms, based on a frequency cut-off of five in the UDHR. We then randomly select two low-frequency words and one high-frequency one, and form a query of the form "(low1 OR low2) AND high", where low1 and low2 are low-frequency words and high is a high-frequency word.

In Section 2 2 we compare these two approaches to query generation, with varying numbers of queries.

Search Engine Queries We used the BootCaT tools to issue the queries generated above to the Bing Web Search API, and retrieve the top-10 result URLs for each query. We further used the BootCaT tools to remove duplicate URLs from the results.

Downloading Content Ferraresi et al. (2008) note that very small documents tend to contain little material appropriate for inclusion in corpora (due to the overhead of HTML markup), whereas very large documents tend to correspond to lists or catalogs, and not more-standard text. Following Ferraresi et al. (2008), we only download documents with size 5–200KB, and MIME-type text/html.

Markup and Boilerplate Removal We remove HTML markup and boilerplate text — e.g., navigation bars, headers, and footers — from the downloaded documents using jusText (Pomikálek, 2011), which preserves paragraph structure present in the HTML in the extracted text. JusText is able to incorporate information from a language-dependent stopword list in determining which document portions are boilerplate and which are more-conventional text. However, in preliminary experiments we observed that using a stopword list derived from the UDHR resulted in many portions of Mi'kmaq text not being recognized as such, and therefore chose to disable this feature.[11]

Language identification Although many language identification tools are available (Cavnar and Trenkle, 1994; Lui and Baldwin, 2012, for example), very little language identification research has considered Mi'kmaq, with Scannell (2007) and Brown (2014) being notable exceptions, and even these have not focused specifically on this language. Here we implement a simple approach to language identification. We represent each language for which the UDHR is available in NLTK as a vector of the character trigram frequencies in the corresponding version of the UDHR. For each document in our corpus, we similarly represent it as a vector of character trigram frequencies, and then compute its cosine similarity with the vector representing each language. We discard any document for which the most-similar language is not Mi'kmaq.

Deduplication The web contains many duplicate and near-duplicate documents. We perform (near and exact)

[6] `https://www.sketchengine.co.uk/`

[7] Scannell (2007) also considers restricted web crawls, using the result URLs as seeds for the crawler, to find additional documents; in this preliminary work we do not consider crawling.

[8] `http://www.un.org/en/universal-declaration-human-rights/`

[9] `http://www.nltk.org/`

[10] `http://bootcat.dipintra.it/`

[11] BTE (Finn et al., 2001) is an alternative boilerplate extraction tool that does not make use of a stopword list, and has been used in many corpus construction efforts (Baroni and Bernardini, 2004; Sharoff, 2006; Ferraresi et al., 2008, for example); however, unlike jusText, BTE does not preserve paragraph structure in the extracted text, and we rely on this structure in subsequent processing.

Queries	Random		Crúbadán	
	# Docs	# Tokens	# Docs	# Tokens
100	36	33k	62	35k
500	88	62k	138	79k
1000	120	92k	167	90k

Table 1: The number of documents and tokens in corpora constructed using the random and Crúbadán approaches to query generation, for increasing numbers of queries.

Corpus	# Docs	# Tokens	# Types
This paper	69	76k	24k
Crúbadán (Mi'kmaq)	31	100k	12k

Table 2: The number of documents, tokens, and types in the corpus created in this paper, and the Mi'kmaq corpus of Crúbadán (Scannell, 2007).

deduplication at the sub-document level based on paragraphs (as provided by jusText), using onion (Pomikálek, 2011), under its default configuration. In this setup, onion iterates through the corpus once, and removes any paragraph for which more than 50% of its 7-grams have been seen in the corpus up to that point.

2.2. Corpus Analysis

We now analyze corpora constructed using the methodology described in the previous subsection, and compare the random and Crúbadán query generation strategies. For this analysis we applied simple whitespace-based tokenization. We further used a preliminary approach to language identification — based on the same methodology as described in Section 2 1 — but that only compared the representation of a given target document to known English, French, and Mi'kmaq text. English and French are official languages of Canada, and preliminary observations indicated that English and French were common in the data prior to language identification.

Table 1 shows the number of documents and tokens in corpora constructed using the random and Crúbadán approaches to query generation, for increasing numbers of queries. The growth in corpus size observed as the number of queries is increased suggests that, in future work, larger corpora could potentially be constructed by issuing more queries. Turning to the differences between the random and Crúbadán approaches to query generation, for each number of queries, the Crúbadán approach gives many more documents than the random approach, while the number of tokens is not drastically different between the two approaches.

Figure 1 shows the number of documents of varying lengths (measured in number of tokens) in the corpora created using the random and Crúbadán query generation strategies with 100 queries. Although the Crúbadán approach gives many more documents, many of these documents are very short. For the remainder of this paper we therefore consider the corpus built using the random approach — which is not dominated by very short documents — using 1000 queries.

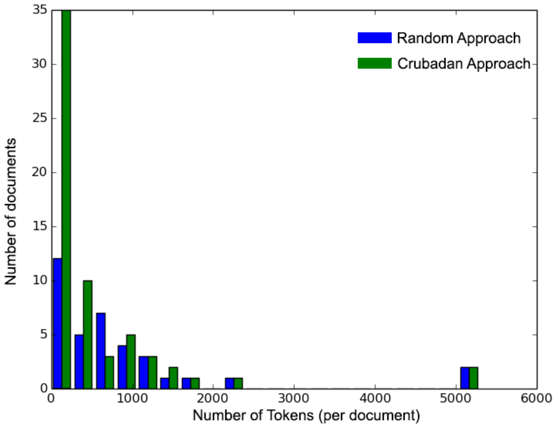

Figure 1: The number of documents, binned by number of tokens per document, for the random and Crúbadán approaches to query generation, using 100 queries.

In Table 2 we compare the number of documents, tokens, and types in the corpus created using the random approach to query generation with 1000 queries ("This paper"), with that of the Crúbadán Mi'kmaq corpus (Scannell, 2007). Here, and for the remainder of the paper, we use the language identification strategy described in Section 2 1 and tokenize our corpus using a simple regular expression-based tokenizer (whereas for the previous analysis a preliminary approach to language identification was used, and tokenization was based on whitespace). At 76k tokens, our corpus is substantially smaller than that of Scannell (2007). These differences in corpus size could be due in part to the relative aggressiveness with which the corpora have been cleaned, in steps such as boilerplate removal and deduplication. Nevertheless, the number of documents and types in our corpus are greater than that of the Crúbadán Mi'kmaq corpus, suggesting that there could be more diversity of authors, text types, or topics in our corpus. In future analysis, quantitative (Kilgarriff, 2001) or qualitative (Kilgarriff, 2012) corpus comparison methods could be applied in an effort to better understand their differences in composition.[12]

A Mi'kmaq speaker analyzed a sample of 25 randomly-selected paragraphs from our corpus to determine the precision of the language identification. 19 paragraphs (76%) were exclusively Mi'kmaq, while 4 paragraphs (16%) contained a mixture of Mi'kmaq and English, for a total of 23 paragraphs (92%) having Mi'kmaq content. Most of the content in the corpus is Mi'kmaq, although there is scope to improve the language identification. In the next section, we use this corpus in language modelling experiments.

3. Language Modelling

Language models are a key component of systems for many NLP tasks. As a first step towards our goal of building NLP systems for Mi'kmaq, we carry out preliminary language

[12] Crucially, this analysis would be possible because the methods of Kilgarriff (2001) and Kilgarriff (2012) require only word frequency lists, which are available for the Crúbadán corpora.

Language model	Bits-per-character
KenLM n=2	2.51
KenLM n=3	2.48
char-rnn	4.36
CNN	2.60

Table 3: Bits-per-character for each approach to language modelling considered.

modelling experiments. Because Mi'kmaq is a polysynthetic language, we consider character-level and subword-level language models, in addition to more-conventional (word) n-gram models. Specifically we consider the following approaches to language modelling:

KenLM We use KenLM (Heafield et al., 2013) to build conventional (word) n-gram language models with modified Kneser-Ney smoothing, for differing orders of n.

char-rnn We consider a character-level LSTM language model. Specifically we use a TensorFlow implementation[13] of char-rnn,[14] with its default parameter settings.

CNN Kim et al. (2016) propose a language model that forms word-level representations, based on a character-level CNN, which then feed into a multi-layer (word-level) LSTM. Kim et al. (2016) found this approach to perform particularly well on morphologically-rich languages, and so we are especially interested in evaluating it on Mi'kmaq. We use a TensorFlow implementation of this model,[15] with the "small" model settings from Kim et al. (2016) because of the relatively small size of our corpus.

We randomly split our corpus into 80% training, 10% development, and 10% testing data, based on sentences. We use a simple regular expression-based approach to detect sentence boundaries. We train our language models on the training data, use the development data for preliminary experiments and parameter tuning, and evaluate our models on the testing data.

KenLM and CNN predict words, and can therefore be evaluated in terms of perplexity. char-rnn, on the other hand, predicts characters, and is evaluated in terms of bits-per-character. Following Hwang and Sung (2017) we convert between perplexity (PPL) and bits-per-character (BPC) as follows:

$$\mathrm{PPL} = 2^{\mathrm{BPC} * \frac{N_C}{N_W}} \tag{1}$$

where N_C and N_W are the number of characters and words, respectively, in the test data.

Table 3 shows bits-per-character for each approach to language modelling considered. In the case of conventional n-gram models, a trigram model (KenLM n=3) outperforms a bigram model (KenLM n=2). Higher order n-gram models (not shown in Table 3) performed comparably to the trigram model. Both neural network-based models that incorporate character-level information, char-rnn and CNN, perform worse than conventional n-gram models. In both cases this could be because sufficient training data is not available to learn the parameters of the neural networks. However, the encouraging performance of CNN suggests that there is also substantial scope for future work to explore this model's various parameter settings — e.g., the dimensionality of the embeddings, the width and number of filters — to better tune it to very low-resource settings, and for the specific case of Mi'kmaq.

4. Discussion

NLP could potentially contribute to the preservation of Indigenous languages — which is particularly important in the current Canadian context of Truth and Reconciliation — by building tools to help use Indigenous languages in computer-mediated communication. To the best of our knowledge this is the first computational work to specifically consider Mi'kmaq. In this preliminary work, we built a web corpus of Mi'kmaq, and evaluated several approaches to language modelling for Mi'kmaq.

Our first direction for future work is to build a larger Mi'kmaq corpus. Our analysis in Section 2 2 indicates that this could be achieved by issuing more search engine queries. On the other hand, Scannell (2007) argues that web crawling using seed URLs returned by the queries is important for building corpora for low-resource languages. Although the corpus created in this paper includes more documents than the Mi'kmaq corpus of Scannell (2007), crawling could still be useful to find more Mi'kmaq documents, and we intend to explore this in future work.

Another important area for future work is language identification. Our analysis in Section 2 2 suggests that other languages are often present along with Mi'kmaq within a single paragraph. Some approaches to language identification are able to recognize which portions of text correspond to a particular language (Jurgens et al., 2017). Extending such methods to recognize Indigenous languages, and Mi'kmaq in particular, is an important area of future work. Future work on Mi'kmaq language identification should also take into account the variation in Mi'kmaq writing systems.

In Section 3 we have identified some potential future directions with respect to language modelling. Recent approaches to low-resource language modelling have incorporated cross-lingual word embeddings learned from bilingual dictionaries (Adams et al., 2017), which are available for Mi'kmaq. We also intend to evaluate such approaches to language modelling in future work.

Finally, because of the morphological complexity of Mi'kmaq, we are particularly interested in morphological analyzers for Mi'kmaq. As a first step, we intend to consider evaluating unsupervised approaches to morphological analysis (Smit et al., 2014) on Mi'kmaq.

[13] https://github.com/crazydonkey200/tensorflow-char-rnn

[14] https://github.com/karpathy/char-rnn

[15] https://github.com/mkroutikov/tf-lstm-char-cnn

5. Acknowledgements

We thank Starlit Simon for her annotation work which enabled the evaluation of the language identification. We thank David Perley for his input on this research. This research is financially supported by Mitacs and the University of New Brunswick.

6. Bibliographical References

Adams, O., Makarucha, A., Neubig, G., Bird, S., and Cohn, T. (2017). Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 937–947, Valencia, Spain.

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the Web. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1313–1316, Lisbon, Portugal.

Baroni, M., Kilgarriff, A., Pomikálek, J., and Rychlý, P. (2006). WebBootCaT: a web tool for instant corpora. In *Proceedings of XII EURALEX International Congress (EURALEX 2006)*, pages 123–131, Torino, Italy.

Battiste, M. (1985). Micmac literacy and cognitive assimilation. In Barbara Burnaby, editor, *Promoting Native Writing Systems in Canada*, pages 7–16. OISE Press/Ontario Institute for Studies in Education, Toronto, Canada.

Bird, S., Loper, E., and Klein, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc., Sebastopol, CA.

Brown, R. (2014). Non-linear mapping for improved identification of 1300+ languages. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 627–632, Doha, Qatar.

Cavnar, W. B. and Trenkle, J. M. (1994). N-gram based text categorization. In *Proceedings of the Third Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, USA.

DeBlois, A. D. (1990). *Micmac Texts*. Canadian Museum of Civilazation, Hull, Canada.

DeBlois, A. D. (1996). *Micmac Dictionary*. Canadian Museum of Civilazation, Hull, Canada.

Ferraresi, A., Zanchetta, E., Baroni, M., and Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4): Can we beat Google?*, pages 47–54, Marrakech, Morocco.

Finn, A., Kushmerick, N., and Smyth, B. (2001). Fact or fiction: Content classification for digital libraries. In *Proceedings of the Second DELOS Network of Excellence Workshop on Personalisation and Recommender Systems in Digital Libraries*, Dublin, Ireland.

Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. (2013). Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, Sofia, Bulgaria.

Hwang, K. and Sung, W. (2017). Character-level language modeling with hierarchical recurrent neural networks. In *Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5720–5724, New Orleans, USA.

Johnson, P. (1996). Mi'kmaq. In Frederick E. Hoxie, editor, *Encyclopedia of North American Indians*, pages 376–378. Houghton Mifflin Company, Boston, USA.

Jurgens, D., Tsvetkov, Y., and Jurafsky, D. (2017). Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 51–57, Vancouver, Canada.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):97–133.

Kilgarriff, A. (2012). Getting to know your corpus. In *Proceedings of Text, Speech, Dialogue (TSD 2012)*, Brno, Czech Republic.

Kim, Y., Jernite, Y., Sontag, D., and Rush, A. M. (2016). Character-aware neural language models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, pages 2741–2749, Phoenix, USA.

Lui, M. and Baldwin, T. (2012). langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012) Demo Session*, pages 25–30, Jeju, Republic of Korea.

Pomikálek, J. (2011). *Removing Boilerplate and Duplicate Content from Web Corpora*. Ph.D. thesis, Masaryk University.

Rand, S. T. (1888). *Dictionary of the language of the Micmac Indians : who reside in Nova Scotia, New Brunswick, Prince Edward Island, Cape Breton and Newfoundland*. Nova Scotia Printing Company, Halifax, Canada.

Scannell, K. (2007). The Crúbadán Project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop, incorporating Cleaneval*, pages 5–15, Louvain-la-Neuve, Belgium.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In Marco Baroni et al., editors, *Wacky! Working papers on the Web as Corpus*, pages 63–98. GEDIT, Bologna, Italy.

Smit, P., Virpioja, S., Grönroos, S.-A., and Kurimo, M. (2014). Morfessor 2.0: Toolkit for statistical morphological segmentation. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 21–24, Gothenburg, Sweden.

Truth and Reconciliation Canada. (2015). *Honouring the truth, reconciling for the future: Summary of the final report of the Truth and Reconciliation Commission of Canada*. Truth and Reconciliation Commission of Canada, Winnipeg, Canada.