# Fine-tuning Sentence-RoBERTa to Construct Word Embeddings for Low-resource Languages from Bilingual Dictionaries

**Diego Bear**
University of New Brunswick
Faculty of Computer Science
diego.bear@unb.ca

**Paul Cook**
University of New Brunswick
Faculty of Computer Science
paul.cook@unb.ca

## Abstract

Conventional approaches to learning word embeddings (Mikolov et al., 2013; Pennington et al., 2014) are limited to relatively few languages with sufficiently large training corpora. To address this limitation, we propose an alternative approach to deriving word embeddings for Wolastoqey and Mi'kmaq that leverages definitions from a bilingual dictionary. More specifically, following Bear and Cook (2022), we experiment with encoding English definitions of Wolastoqey and Mi'kmaq words into vector representations using English sequence representation models. For this, we consider using and fine-tuning sentence-RoBERTa models (Reimers and Gurevych, 2019). We evaluate our word embeddings using a similar methodology to that of Bear and Cook using evaluations based on word classification, clustering and reverse dictionary search. We additionally construct word embeddings for higher-resource languages — English, German and Spanish — using our methods and evaluate our embeddings on existing word-similarity datasets. Our findings indicate that our word embedding methods can be used to produce meaningful vector representations for low-resource languages such as Wolastoqey and Mi'kmaq and for higher-resource languages.

## 1 Introduction

Word embeddings (Mikolov et al., 2013; Pennington et al., 2014) are real-numbered vector representations of the meanings of words and are a fundamental component of many natural language processing (NLP) systems. Although word embeddings can often be learnt while training NLP systems end-to-end, pretrained word embeddings have been shown to bolster the performance of NLP systems in tasks such as machine translation (Qi et al., 2018) and information retrieval (Roy et al., 2018). Despite their utility, quality word embeddings can be difficult to obtain as they generally require large corpora of running text to train. This represents a significant limitation of conventional word embedding methods as, due to these data requirements, quality word embeddings can only be learnt for relatively few languages. Today, a majority of languages spoken around the world are low-resource (Arppe et al., 2016), and thus lack the text resources required to train high quality word embeddings. As this is the case, an alternative embedding approach is desirable to make better use of what data exists for low-resource languages.

In the case of Wolastoqey (also referred to as Passamaquoddy-Maliseet) and Mi'kmaq, there simply isn't enough data available in these languages to train quality word embeddings using conventional methods. Wolastoqey and Mi'kmaq are both low-resource Eastern Algonquin languages. There are currently approximately 300 remaining first language speakers of Wolastoqey and $7k$ speakers of Mi'kmaq (Statistics Canada, 2017) in Canada. Due to the low-resource state of these languages, developing language technologies for Wolastoqey and Mi'kmaq is challenging because there are no large corpora or annotated datasets available in these languages to train NLP systems. Despite not having large corpora or datasets available, both a bilingual Wolastoqey–English dictionary, known as the Passamaquoddy-Maliseet Dictionary (Francis and Leavitt, 2008), and a bilingual Mi'kmaq-English dictionary, known as the Mi'gmaq/Mi'kmaq Online Dictionary (Haberlin et al., 1997), are available. These dictionaries contain English definitions for Wolastoqey and Mi'kmaq headwords and consist of a total of $18.6k$ and $6.5k$ entries, respectively. In our work, we experiment with using these dictionaries to construct word embeddings for Wolastoqey and Mi'kmaq.

Previous work has demonstrated that bilingual lexicons and monolingual corpora can be leveraged to train cross-lingual word embeddings for low-resource languages. For example, Adams et al. (2017) showed that, by combining a large English



*Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 47–57

corpus with a small Yongning Na corpus, and by replacing words with their translations using a small bilingual lexicon, a pseudo-bilingual corpus can be created which can be used to train cross-lingual word embeddings. We do not consider this approach in our work because Wolastoqey and Mi'kmaq are polysynthetic languages, and as such, many tokens that occur in a corpus would not be expected to be found as dictionary headwords, which limits the applicability of this approach.

Instead, we look towards approaches based on sequence representation. Prior work has demonstrated that, by leveraging bilingual dictionaries, useful vector representations can be constructed for Nêhiyawêwin (Plains Cree) words. By averaging word embeddings corresponding to words that appear in English definitions of Nêhiyawêwin words, embeddings can be obtained which can be used to effectively cluster Nêhiyawêwin words (Harrigan and Arppe, 2021) and map them to preconstructed ontologies (Dacanay et al., 2021).

Bear and Cook (2022) extended the methodology of Harrigan and Arppe (2021) and Dacanay et al. (2021) to construct word embeddings for Wolastoqey. They used the average of word2vec embeddings to represent words from their dictionary definitions, as well as RoBERTa, and sentence-RoBERTa models to encode definitions into vector representations. These embeddings were then evaluated based on word classification tasks focused on predicting part-of speech, animacy, and transitivity; semantic clustering; and reverse dictionary search. In each evaluation, it was found that approaches using these embeddings outperformed task-specific baselines, indicating that sentence-transformer models can outperform approaches based on word embeddings for this purpose.

As this approach has been shown to perform relatively well, in this paper, we build upon the work of Bear and Cook (2022) by fine-tuning sequence representation models for this task. More specifically, we propose fine-tuning sentence-RoBERTa models on monolingual dictionary definitions to determine if doing so could improve the overall quality of the representations. Using these fine-tuned models, we construct word embeddings for Wolastoqey and Mi'kmaq from English definitions in the Passamaquoddy-Maliseet Dictionary and Mi'gmaq/Mi'kmaq Online Dictionary.

Following Bear and Cook (2022), we evaluate our Wolastoqey and Mi'kmaq word embeddings on word classification tasks focused on predicting part-of speech, animacy, and transitivity as well as semantic clustering and reverse dictionary search. We compare our word embeddings against task-specific baselines and embeddings produced using the techniques of Bear and Cook. To assess if this technique is viable for other higher-resource languages, we also construct word embeddings for English, Spanish and German, and evaluate the performance of our models on word similiarity datasets, comparing against previously reported results.

## 2 Methodology

To obtain embeddings for Wolastoqey and Mi'kmaq words, we experiment with encoding English definitions of Wolastoqey words in the Passamaquoddy-Maliseet Dictionary, and Mi'kmaq words in the Mi'gmaq/Mi'kmaq Online Dictionary, into vector representations. To construct vector representations from English definitions, we consider fine-tuning and using sentence-transformer models, masked language models specifically trained for sequence representation. More specifically, we consider fine-tuning sentence-RoBERTa models using three training regimens from Reimers and Gurevych (2019). We compare our embeddings constructed with this approach to those created using the methodology of Bear and Cook (2022).

In our work, we fine-tune our sentence-RoBERTa models on the dataset of Hill et al. (2016). This dataset consists of word–definition pairs collected from several English dictionaries and WordNet (Miller, 1995). In our experiments, we use the training and development splits from Zheng et al. (2020). This gives us a training set consisting of a total of $667.5k$ word–definition pairs corresponding to $45k$ unique types and a development set consisting of $75.8k$ definitions corresponding to $5k$ unique types. However, due to the data requirements of our fine-tuning regimens, we filter out any definitions corresponding to words with only one unique definition and filter out duplicate definitions from both our training and development sets. This reduces our effective training corpus size to $664.7k$ definitions corresponding to $42.5k$ unique types, and our development set to $75.6k$ definitions corresponding to $4.7k$ unique types. We use our development set to ensure overfitting does not occur and to monitor training performance.

To fine-tune our sentence-RoBERTa models,

we continue training from the nli-roberta-base-v2 model available in the sentence transformer 2.1.0 library.[1] This model represents a checkpoint that has been pretrained on a large natural language inference dataset, constructed by combining the Stanford NLI corpus (Bowman et al., 2015) and the multi-genre NLI corpus (Williams et al., 2018). We consider fine-tuning three models using the softmax, cosine, and triplet loss training objectives outlined in Reimers and Gurevych (2019). Each of these training objectives requires the model to be trained in a Siamese configuration in which two or more examples are passed through the network independently before being compared to compute training loss at a given time-step.

The softmax training objective is based on classification. In our work, the classification task we fine-tune our model on is determining if two definitions correspond to the same word. To construct training pairs for this fine-tuning regimen, we pair each definition in our training set with another definition to form either a positive or negative training example. We assign half of our definitions another definition corresponding to the same word, forming a positive pair, and we assign the other half definitions that do not correspond to the same word, forming negative pairs. This gives us 664.7$k$ training examples, equal to the number of definitions in our training corpus.

The cosine training objective is based on regression. More specifically, in this fine-tuning regimen, we attempt to match the cosine similarity between two output vectors to some ground truth label. To obtain examples, we form training pairs similarly to how we did for the softmax fine-tuning regimen. However, instead of assigning a binary label to pairs, we assign a ground truth cosine similarity. For positive pairs, this is simply equal to 1.0. However, for negative pairs, to obtain ground-truth cosine similarities, we use the cosine similarities computed from vectors in a word2vec model. For this purpose, we use a word embedding model that has been trained on a Google News corpus consisting of roughly 100 billion words.[2] We obtain these embeddings using gensim 3.8.3 (Řehůřek and Sojka, 2010). For each negative sample, the ground-truth cosine similarity used for training is set to the cosine similarity calculated using the embeddings for the words each definition corresponds to. In this training configuration, loss is calculated as the mean squared error between the cosine similarity of the two input vectors and the ground truth reference.

Finally, triplet loss considers three inputs, in our case definitions, at a given timestep. More specifically, this training scheme requires an anchor, as well as two additional inputs that act as positive and negative instances. When fine-tuning with this training objective, we attempt to learn weights such that the representations produced for each anchor are closer to their corresponding positive than negative instance. As this is the case, to form training examples, we treat each definition in our training set as an anchor and assign each an accompanying positive instance — a definition corresponding to the same word — and a negative instance — a definition corresponding to a different word. Like before, this gives us a total of 664.7$k$ training examples to fine-tune our model with.

For each training technique considered, we fine-tune our models using the default training parameters of the sentence-transformers library. We fine-tune our models for a single epoch, as, training for three epochs appeared to degrade performance on our word classification tasks in early testing. After fine-tuning, we are left with three models, each fine-tuned using a different training regimen.

To construct word embeddings using these models, we first preprocess our input definitions using the same preprocessing steps as Bear and Cook (2022). Namely, we consider removing bracketed content from our input definitions, as, in the dictionaries we use in our work, this typically consists of topical information that does not contribute to the core meaning of definitions. We then pass our preprocessed input definitions to our sentence-RoBERTa models to obtain a vector representation based on the mean output vectors of our sentence-RoBERTa models.

## 3 Word Classification

Following Bear and Cook (2022), we evaluate our word embeddings on word classification tasks to determine if they are capable of capturing information about the syntactic properties of words. We consider three word classification tasks focused on predicting, 1.) part-of-speech, 2.) noun animacy and 3.) verb type. For each task, we train logistic regression classifiers to predict the syntactic labels of words from their embeddings.

[1] https://www.sbert.net/

[2] https://code.google.com/archive/p/word2vec/

### 3.1 Experimental Setup

To construct datasets for these evaluations, we use gold-standard labels from the Passamaquoddy-Maliseet Dictionary and Mi'gmaq/Mi'kmaq Online Dictionary. For our part-of-speech classification tasks, we consider a total of $18k$ entries from the Passamaquoddy-Maliseet Dictionary, consisting of 53 pronouns, 231 preverbs, 570 particles, $13.7k$ verbs and $3.3k$ nouns, for Wolastoqey and $6.4k$ entries from the Mi'gmaq/Mi'kmaq Online Dictionary, consisting of 16 pronouns, 119 particles, $4.6k$ verbs and $1.6k$ nouns, for Mi'kmaq. For our noun animacy classification tasks, we remove any entries corresponding to words that can occur as both animate and inanimate. In total, we use $1.7k$ animate, and $1.3k$ inanimate nouns for Wolastoqey and 756 animate, and 806 inanimate, nouns for Mi'kmaq.

In both Wolastoqey and Mi'kmaq, verbs are categorized into four distinct groups based on their combination of animacy and transitivity. More specifically, Wolastoqey and Mi'kmaq verbs can be, animate intransitive, inanimate intransitive, transitive animate, or transitive inanimate. We remove any entries that correspond to more than one of these categories. This gives a total of $5.3k$ animate intransitive, $2.1k$ inanimate intransitive, $3k$ transitive animate, and $2.7k$ transitive inanimate Wolastoqey verbs, and $2k$ animate intransitive, 753 inanimate intransitive, $1k$ transitive animate and 847 transitive inanimate Mi'kmaq verbs, for our verb type classification tasks.

To conduct this evaluation, we first construct embeddings for each Wolastoqey and Mi'kmaq word using our proposed methodology. We then train logistic regression classifiers for each task and method. For this evaluation, we implement our logistic regression classifiers using scikit-learn 0.24.2. We use the default training parameters of this library, except max-iterations, which we set to 6000, so that all models finish converging. We train and evaluate in a 10-fold cross validation setup. We use macro-averaged accuracy, precision, recall, and F1-score as our evaluation metrics and compare our models to a most-frequent class baseline as well as the pretrained sentence-RoBERTa based approach proposed by Bear and Cook (2022) as it has been shown to achieve strong performance in this task.

### 3.2 Results

Results are shown in Table 1 for Wolastoqey and Table 2 for Mi'kmaq. We observe that, for all

| Part of Speech | | | | |
|---|---|---|---|---|
| Method | Accuracy | P | R | F1 |
| Most Freq. | 0.767 | 0.153 | 0.200 | 0.174 |
| sRoBERTa | 0.974 | 0.828 | 0.801 | 0.809 |
| Cosine | 0.976 | 0.858 | **0.829** | **0.839** |
| Softmax | 0.976 | 0.855 | **0.829** | 0.838 |
| Triplet | **0.979** | **0.862** | 0.823 | 0.837 |
| Noun Animacy | | | | |
| Most Freq. | 0.552 | 0.276 | 0.500 | 0.355 |
| sRoBERTa | 0.801 | 0.800 | 0.798 | 0.798 |
| Cosine | 0.804 | 0.804 | 0.804 | 0.803 |
| Softmax | 0.789 | 0.791 | 0.787 | 0.787 |
| Triplet | **0.806** | **0.805** | **0.805** | **0.804** |
| Verb Type | | | | |
| Most Freq. | 0.406 | 0.101 | 0.250 | 0.144 |
| sRoBERTa | **0.951** | **0.953** | **0.953** | **0.953** |
| Cosine | 0.921 | 0.926 | 0.925 | 0.925 |
| Softmax | 0.932 | 0.936 | 0.934 | 0.935 |
| Triplet | 0.947 | 0.950 | 0.950 | 0.950 |

Table 1: Results for each Wolastoqey word classification task using each embedding method, and a most-frequent class baseline. The best result for each task and metric is shown in boldface.

| Part of Speech | | | | |
|---|---|---|---|---|
| Method | Accuracy | P | R | F1 |
| Most Freq. | 0.730 | 0.182 | 0.250 | 0.211 |
| sRoBERTa | 0.973 | 0.823 | 0.795 | 0.800 |
| Cosine | 0.973 | 0.847 | 0.839 | 0.834 |
| Softmax | 0.976 | 0.844 | 0.819 | 0.823 |
| Triplet | **0.977** | **0.861** | **0.841** | **0.841** |
| Noun Animacy | | | | |
| Most Freq. | 0.516 | 0.258 | 0.500 | 0.340 |
| sRoBERTa | 0.764 | 0.766 | 0.764 | 0.763 |
| Cosine | 0.783 | **0.786** | 0.782 | 0.782 |
| Softmax | 0.777 | 0.777 | 0.776 | 0.776 |
| Triplet | **0.784** | 0.785 | **0.784** | **0.783** |
| Verb Type | | | | |
| Most Freq. | 0.439 | 0.110 | 0.250 | 0.152 |
| sRoBERTa | 0.865 | 0.861 | 0.860 | 0.860 |
| Cosine | 0.845 | 0.843 | 0.840 | 0.840 |
| Softmax | 0.850 | 0.849 | 0.845 | 0.846 |
| Triplet | **0.872** | **0.868** | **0.868** | **0.867** |

Table 2: Results for each Mi'kmaq word classification task using each embedding method, and a most-frequent class baseline. The best result for each evaluation metric and task is shown in boldface.

tasks and evaluation metrics, all of our models outperform a most-frequent class baseline. This indicates that these approaches to representing Wolastoqey and Mi'kmaq words capture information about these syntactic properties.

We observe fine-tuning sentence-RoBERTa on English dictionary definitions leads to improved performance on classification tasks involving Wolastoqey nouns, however, it decreases performance on our Wolastoqey verb classification task. Of our fine-tuned sentence-RoBERTa models, the model trained with the triplet training objective performs the best on each Wolastoqey task except part-of-speech classification.

For Mi'kmaq, we again see that fine-tuning sentence-RoBERTa with our cosine and softmax training objectives results in a decrease in performance on verb classification but increases performance on part-of-speech and noun animacy classification. However, here we observe that our sentence-RoBERTa model fine-tuned with triplet loss outperforms all other models considered in all classification tasks in terms of accuracy and F1 score. From these results, and the results from our Wolastoqey evaluation, of our fine-tuned models, the model trained with our triplet loss is best able to represent Wolastoqey and Mi'kmaq words from their definitions.

## 4 Clustering

Here we explore using our embedding models to semantically cluster Wolastoqey and Mi'kmaq words. For this experiment, we largely follow the evaluation procedures of Bear and Cook (2022). For Wolastoqey, we reproduce the experiments of Bear and Cook for the purpose of comparison.

### 4.1 Experimental Setup

To perform our clustering evaluations, we require ground-truth labels to compare our results to. In the case of Wolastoqey, we consider using the same dataset as Bear and Cook (2022) for this purpose. More specifically, we consider obtaining categorical labels from Wolastoqewatu,[3] a website designed to help teach Wolastoqey, and Wolastoqey Latuwewakon,[4] a mobile application designed to teach Wolastoqey vocabulary. For Wolastoqewatu, we use the glossary categories as labels, while for Wolastoqey Latuwewakon, we use the top-level categories from the categories tab. We filter out words that appear in multiple categories and cross-reference the remaining words with the Passamaquoddy-Maliseet Dictionary to obtain word–category pairs. In total, using this approach, we are left with 1154 entries from Wolastoqewatu that correspond to 20 unique categories and 78 entries from Wolastoqey Latuwewakon that correspond to 6 unique categories.

To obtain gold-standard labels for our Mi'kmaq clustering evaluation, we use categories from the Mi'gmaq/Mi'kmaq Online Dictionary, which contains a glossary consisting of words grouped into topically-organized categories. We use these categories as ground truth references for our clustering evaluation. Using these labels, we create an evaluation set consisting of 6465 items corresponding to 237 classes. However, unlike our aforementioned Wolastoqey datasets, words in this evaluation set frequently correspond to more than one class. As this is the case, we do not remove these words from the evaluation set.

To cluster the words in each dataset, we use K-means, setting the number of clusters to the number of classes in each dataset (i.e., 20 for Wolastoqewatu, 6 for Wolastoqey Latuwewakon, and 237 for the Mi'kmaq dictionary dataset). For this, we use the default parameters of the scikit-learn 0.24.2 implementation of K-means. We evaluate the clustering using BCubed precision, recall, and F1-score. We compare our proposed methods to the pretrained sentence-RoBERTa approach of Bear and Cook (2022) to determine if our fine-tuning proceedures improve over pretrained sentence-RoBERTa models for this task.

### 4.2 Results

Results are shown in Table 3. We observe that additionally fine-tuning sentence-RoBERTa on monolingual dictionary definitions results in mixed improvements. On the Wolastoqewatu dataset, the only model that substantially outperforms our pretrained sentence-RoBERTa model across metrics is the softmax model. However, this does not hold true for the Wolastoqey Latuwewakon dataset, where all models fine-tuned using monolingual dictionary definitions outperform the pretrained sentence-RoBERTa model in terms of BCubed F1 score.

We observe different trends on our Mi'kmaq evaluation. Here, we observe that our pretrained

[3] https://wolastoqewatu.ca

[4] https://wolastoqey-latuwewakon.web.app/

| Wolastoqewatu | | | |
|---|---|---|---|
| Method | BCubed P | BCubed R | BCubed F1 |
| s-RoBERTa | 0.371 | 0.324 | 0.346 |
| Cosine | 0.348 | 0.296 | 0.320 |
| Softmax | 0.392 | **0.334** | **0.360** |
| Triplet | **0.391** | 0.316 | 0.349 |
| Wolastoqey Latuwewakon | | | |
| Method | BCubed P | BCubed R | BCubed F1 |
| s-RoBERTa | 0.668 | 0.496 | 0.569 |
| Cosine | 0.706 | 0.546 | 0.615 |
| Softmax | **0.732** | **0.553** | **0.630** |
| Triplet | 0.722 | 0.515 | 0.601 |
| Mi'gmaq/Mi'kmaq Online Dictionary | | | |
| Method | BCubed P | BCubed R | BCubed F1 |
| s-RoBERTa | **0.347** | **0.122** | **0.181** |
| Cosine | 0.259 | 0.080 | 0.122 |
| Softmax | 0.329 | 0.108 | 0.162 |
| Triplet | 0.343 | 0.113 | 0.170 |

Table 3: Clustering evaluation results for each embedding method on each dataset. The best result for each evaluation metric and dataset is shown in boldface.

sentence-RoBERTa model substantially outperforms all other models in each evaluation metric, and that fine-tuning sentence-RoBERTa results in worse performance on all metrics.

Unlike the Wolastoqewatu and Wolastoqey Latuwewakon datasets, which consist mostly of nouns, the Mi'kmaq dataset is primarily composed of verbs. This could be why we see different trends in the results on this dataset. The finding that pretrained sentence-RoBERTa outperforms our fine-tuned models on our Mi'kmaq evaluation is consistent with the findings from 3.2 that our fine-tuned cosine and softmax models generally performed better than pretrained sentence-RoBERTa on Mi'kmaq classification tasks involving nouns, but, worse than pretrained sentence-RoBERTa on our verb classification task (Table 2). The findings for our triplet model, which performed slightly better on Mi'kmaq verb classification experiments than our pretrained sentence-RoBERTa model, are, however, not consistent with this.

## 5 Reverse Dictionary

Here we use our Wolastoqey and Mi'kmaq word representations to create reverse dictionary search systems. Such systems could potentially help Wolastoqey and Mi'kmaq learners to more-easily access language resources.

### 5.1 Datasets

We build datasets for our reverse dictionary search evaluations based on the principle that the English definition for a Wolastoqey word in the Passamaquoddy-Maliseet Dictionary, or a Mi'kmaq word in the Mi'gmaq/Mi'kmaq Online Dictionary, is expected to be similar to an alternative English definition for that word from another dictionary. In this evaluation, we use alternative English definitions for Wolastoqey and Mi'kmaq words as simulated queries, which we compare against search spaces composed of reference definitions from the Passamaquoddy-Maliseet Dictionary and Mi'gmaq/Mi'kmaq Online Dictionary.

As there are relatively few data sources containing English definitions for Wolastoqey and Mi'kmaq words, we use a similar approach to Bear and Cook (2022) to obtain alternative definitions for the Wolastoqey and Mi'kmaq words in our search spaces. We leverage the fact that many definitions in the Passamaquoddy-Maliseet Dictionary and Mi'gmaq/Mi'kmaq Online Dictionary are composed of a single-word. More specifically, we use an English dictionary, namely WordNet (Miller, 1995), to find alternative definitions for each Wolastoqey and Mi'kmaq word corresponding to a single-word definition in the Passamaquoddy-Maliseet Dictionary and Mi'gmaq/Mi'kmaq Online Dictionary. For each single-word definition in the Passamaquoddy-Maliseet Dictionary and Mi'gmaq/Mi'kmaq Online Dictionary that also occurs as a lemma in WordNet, we use the definition for the first WordNet synset associated with that lemma as a simulated query for this evaluation. Using the Wolastoqey word *amalhihpuwakon*, defined as 'dessert' in the Passamaquoddy-Maliseet Dictionary, as an example, we would use the WordNet definition 'a dish served as the last course of a meal' as an alternative definition for this word.

To expand the number of simulated queries available for our evaluations, we also use this approach to obtain alternative definitions for words that correspond to definitions that become single-words after certain words are removed. As dependent nouns and verbs are given in a third person form in the Passamaquoddy-Maliseet Dictionary and the Mi'gmaq/Mi'kmaq Online Dictionary, to obtain alternative definitions for these words, when identifying single word definitions, we remove the words *s/he* and *h/* (abbreviations for *she/he* and *her/his*, respectively) as well as *it* from definitions in the Passamaquoddy-Maliseet Dictionary and all instances of *he/she*, *him/her*, *it*, and *him/her/it* from definitions in the Mi'gmaq/Mi'kmaq Online Dictionary.

As both dictionaries contain definitions for a number of names, we remove all dictionary headwords corresponding to English names from both the pool of single-word definitions, as well as our search spaces, using a list of English names obtained from NLTK (Bird et al., 2009). In total, our approach gives 1091 Wolastoqey words and 1424 words from the Mi'gmaq/Mi'kmaq Online Dictionary, with alternative English definitions available in WordNet. We compare these alternative definitions to search spaces consisting of $17.9k$ Wolastoqey words and $6.4k$ Mi'kmaq words obtained from the Passamaquoddy-Maliseet Dictionary and the Mi'gmaq/Mi'kmaq Online Dictionary respectively.

### 5.2 Experimental Setup

To perform our reverse dictionary search evaluations, we construct vector representations for both the definitions in our search spaces and our simulated queries using our proposed embedding approaches. Using these vector representations, we calculate the cosine distances between each simulated query and each definition in its corresponding search space. We then use the resulting rank of the word corresponding to the simulated queries to calculate our evaluation metrics. Specifically, we consider median rank, mean reciprocal rank (MRR), and accuracy@$k$, for $k = 1, 5, 10, 20, 50, 100$.

### 5.3 Results

Results are shown in Table 4. Of our fine-tuned models, we observe that the model trained with the triplet loss training objective performs best, substantially improving over both the cosine and softmax models in terms of median rank and MRR for both languages. This model also outperforms pretrained sentence-RoBERTa for each evaluation metric and language, except for median rank on Mi'kmaq.

Despite all models outperforming the random baseline, the findings for our best model, the sentence-RoBERTa model fine-tuned using triplet loss, do not suggest that this could yet be used as a practical reverse dictionary search system. For example, the accuracy@100 of 0.544 for Wolastoqey indicates that only roughly half the time is this approach able to rank the correct word among the top-100. The disparity in length and complexity between our query definitions from WordNet and the single-word definitions from the Passamaquoddy-Maliseet Dictionary, and the Mi'gmaq/Mi'kmaq Online Dictionary, used in this evaluation could contribute towards making this experimental setup a particularly challenging task.

## 6 Word Similarity

Although our primary interest is methods for learning Wolastoqey and Mi'kmaq word representations, here we consider whether the proposed approach to encoding dictionary definitions can also be applied to represent words in higher-resource languages. Word similarity datasets are available for many languages and are commonly used to evaluate how well word embedding models are able to capture the similarity or relatedness between words. Here we consider constructing word embeddings for English, German and Spanish using our proposed methodologies and evaluating on word similarity datasets. As these datasets are frequently used in other works, where available, we compare against previously reported results for word2vec baselines.

### 6.1 Experimental Setup

In our experiments, we choose to use one Spanish, one German and two English word similarity datasets. For English, we consider SimLex-999 (Hill et al., 2015) as well as the MEN dataset (Bruni et al., 2014). We use these datasets, as SimLex-999 reflects word similarity, whereas the MEN dataset reflects relatedness. For the MEN dataset, we consider using the full 3000 word pair version of this dataset in our evaluation. For Spanish, we consider using a translation of WordSim-353 (ES-WS353, Finkelstein et al., 2002; Hassan and Mihalcea, 2009) and we use GUR350 (Gurevych, 2005) for German.

We construct embeddings for the words in these datasets using the same approach used to obtain word embeddings for Wolastoqey and Mi'kmaq in our prior evaluations. However, here we do not remove bracketed text from definitions, a pre-processing step motivated specifically based on common patterns in definitions of the Passamaquoddy-Maliseet Dictionary. To construct our English word embeddings, we consider using dictionary definitions from WordNet. As our method requires English definitions for non-English words, for words in the Spanish and German evaluation sets, we construct embeddings using web-scraped definitions from the Collins Spanish–English and German–English online dictionaries (HarperCollins, 2011).

| Wolastoqey Search Space | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Method | Median | MRR | Acc@1 | Acc@5 | Acc@10 | Acc@20 | Acc@50 | Acc@100 |
| Random | 9164 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 | 0.005 |
| Bear and Cook (2022) | 107 | 0.081 | 0.027 | 0.128 | 0.183 | 0.260 | 0.397 | 0.495 |
| Cosine | 311 | 0.056 | 0.025 | 0.072 | 0.118 | 0.170 | 0.269 | 0.350 |
| Softmax | 87 | 0.098 | 0.044 | 0.140 | 0.213 | 0.302 | 0.412 | 0.518 |
| Triplet | **70** | **0.109** | **0.050** | **0.155** | **0.239** | **0.332** | **0.448** | **0.544** |
| Mi'kmaq Search Space | | | | | | | | |
| Method | Median | MRR | Acc@1 | Acc@5 | Acc@10 | Acc@20 | Acc@50 | Acc@100 |
| Random | 3300 | 0.001 | 0.000 | 0.000 | 0.000 | 0.002 | 0.006 | 0.015 |
| Bear and Cook (2022) | **27** | 0.174 | 0.086 | 0.263 | 0.364 | 0.464 | 0.568 | 0.634 |
| Cosine | 108 | 0.111 | 0.060 | 0.148 | 0.215 | 0.301 | 0.409 | 0.493 |
| Softmax | 37 | 0.181 | 0.099 | 0.261 | 0.343 | 0.435 | 0.545 | 0.633 |
| Triplet | 28 | **0.198** | **0.107** | **0.296** | **0.386** | **0.466** | **0.581** | **0.667** |

Table 4: Median rank, MRR, and accuracy@$k$ for each threshold considered, for reverse dictionary experiments using each approach to representing Wolastoqey and Mi'kmaq words and a random baseline.

As it is expected that a number of words will not have definitions in the dictionaries we use, we set the embedding of any word without a definition to a vector of zeroes.

To evaluate how well our embeddings perform, we calculate cosine similarities for word pairs in each dataset using our embedding models. We then calculate the Spearman correlation between the predicted cosine similarities and the human annotated similarity scores for each dataset.

To establish a baseline, for SimLex-999, ES-WS353, and GUR350, we compare our models to previously reported results. More specifically, for SimLex-999, we compare our models to the word2vec results published by (Hill et al., 2015). For ES-WS353 and GUR350, we compare our models to the skipgram results reported in Bojanowski et al. (2017). For the MEN dataset, we calculate a baseline for comparison directly using a word2vec model. Here we use the same Google-News word2vec embeddings as in Section 2. As many words in the MEN dataset use British English spelling, and this word2vec model uses primarily American English spelling, we convert any British English word-forms not found in this embedding model to their American English equivalent.

### 6.2 Results

Results are shown in Table 5. We observe that on all datasets, except SimLex-999, our proposed embedding approaches do not outperform the chosen word2vec baselines. Despite this, all our models, achieve statistically significant correlation on all word similarity datasets considered. We observe that pretrained sentence-RoBERTa outperforms a word2vec baseline on SimLex-999, but fails to do so on the MEN dataset. This could indicate that these embeddings better capture word similarity than relatedness.

| Method | Simlex-999 | MEN | ES-WS353 | GUR350 |
|---|---|---|---|---|
| Baseline | 0.414 | **0.78** | **0.57** | **0.61** |
| sRoBERTa | **0.423** | 0.568 | 0.297 | 0.538 |
| Cosine | 0.374 | 0.524 | 0.313 | 0.494 |
| Softmax | 0.416 | 0.560 | 0.334 | 0.579 |
| Triplet | 0.420 | 0.560 | 0.303 | 0.506 |

Table 5: Spearman correlations between cosine similarities and human-annotated similarity scores for each method on each dataset. The best correlation for each dataset is shown in boldface.

Further fine-tuning sentence-RoBERTa does not improve performance on either English dataset. Despite this, all fine-tuned models outperform pretrained sentence-RoBERTa on ES-WS353 and our softmax model outperforms pretrained sentence-RoBERTa model on GUR350. Definition length may be a factor here, as the pre-trained sentence-RoBERTa model performs best on our English datasets, in which words have an average definition length of 11 tokens, whereas words in our Spanish and German datasets have an average definition length of 1 and 2 tokens, respectively. This would be consistent with the findings from Table 2, in which our fine-tuned models performed better in terms of F1-score on Mi'kmaq classification tasks involving nouns, which have comparatively short definitions, and worse on tasks involving verbs which tend to have longer definitions. However, definition length alone isn't enough to explain the disparity in model rankings, as, in contrast to the results observed for Mi'kmaq, the softmax model failed to outperform pretrained sentence-RoBERTa in Wolastoqey noun animacy classification. As this is the case, the best model configuration seems

to be dependant on the language and task being considered.

## 7 Conclusions

In this paper, we considered approaches to forming word embeddings for Wolastoqey and Mi'kmaq based on their English definitions in bilingual dictionaries. Specifically we considered approaches to fine-tuning sentence-RoBERTa for this. Our findings indicate that our proposed approaches can be used to construct embeddings for Wolastoqey and Mi'kmaq words that capture syntactic and semantic information, and that fine-tuning often gives improvements over pre-trained sentence-RoBERTa, although this improvement is not consistent across languages, tasks, and approaches to fine-tuning. Our results from reverse dictionary evaluations indicate that these embeddings cannot yet be used to build a practical reverse dictionary search system. We further showed that these approaches can be applied to form embeddings for higher-resource languages. Although here these embeddings achieved significant correlations on word similarity and relatedness evaluations, they did not improve over conventional word2vec embeddings.

In future work, we intend to explore ways to improve the embeddings. Although we observed that fine-tuning sentence-RoBERTa did not give consistent improvements across tasks, we hypothesize that an alternative approach could give improvements. Definitions for verbs in the Passamaquoddy-Maliseet Dictionary in particular tend to be longer, while definitions for nouns are typically quite short and often composed of only a single word. This disparity in definition complexity could hinder the effectiveness of our proposed word embedding techniques. We therefore intend to explore the use of ULR-BERT (Li and Zhao, 2021), which is capable of representing words, phrases and sentences proficiently, for forming improved embeddings for Wolastoqey and Mi'kmaq words from their English definitions in bilingual dictionaries.

In addition to using ULR-BERT, we also intend to fine-tune sentence-transformer models that make use of different network architectures and pretraining regimens. In our work, we use a single RoBERTa checkpoint, pretrained on natural language inference, as a uniform starting point for fine-tuning. However, since the release of the original work on sentence transformers, other models have been made available through the sentence-BERT library, for example models based on MPNet (Song et al., 2020), which have been shown to outperform sentence-RoBERTa on sentence embedding benchmarks. As this is the case, the use of these models in-place of sentence-RoBERTa could potentially improve the quality of word embeddings produced using our methodology.

In our work, we demonstrated that we can construct meaningful word embeddings for Wolastoqey and Mi'kmaq dictionary headwords. In future work we will consider evaluating the impact of these embeddings on down-stream applications.

## Limitations

Although improving the performance of our embedding methods is desirable, the most apparent limitation of our work is not the overall quality of representations produced, but rather the range of words our methodologies can be applied to. Currently, our methodology can only be used to construct word embeddings for dictionary headwords. This represents a considerable limitation, as Wolastoqey and Mi'kmaq are both polysyntheic languages, in which speakers often build new words by creatively combining roots. As this is the case, no dictionary is expected to contain definitions for all word-forms of these languages. Because of this, future work is required to extend our approach to construct embeddings for words that do not appear in a bilingual dictionary.

## References


Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 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. Association for Computational Linguistics.

Antti Arppe, Jordan Lachler, Trond Trosterud, Lene Antonsen, and Sjur N Moshagen. 2016. Basic language resource kits for endangered languages: A case study of plains cree. In *Proceedings of the 2nd Workshop on Collaboration and Computing for Under-Resourced Languages Workshop (CCURL 2016), Portorož, Slovenia*, pages 1–8.

Diego Bear and Paul Cook. 2022. Leveraging a bilingual dictionary to learn wolastoqey word representations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1159–1166, Marseille, France. European Language Resources Association.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Int. Res.*, 49(1):1–47.

Daniel Dacanay, Atticus Harrigan, Arok Wolvengrey, and Antti Arppe. 2021. The more detail, the better? – investigating the effects of semantic ontology specificity on vector semantic classification with a Plains Cree / nêhiyawêwin dictionary. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 143–152, Online. Association for Computational Linguistics.

Lev Finkelstein, Evgeniy Gabrilovich, Y. Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: the concept revisited. *ACM Trans. Inf. Syst.*, 20:116–131.

David A. Francis and Robert M. Leavitt. 2008. A passamaquoddy-maliseet dictionary.

Iryna Gurevych. 2005. Using the structure of a conceptual network in computing semantic relatedness. In *Natural Language Processing – IJCNLP 2005*, pages 767–778.

Sean Haberlin, Eunice Metallic, Diane Mitchell, Watson Williams, Joe Wilmot, and Dave Ziegler. 1997. [link].

HarperCollins. 2011. Collins english dictionary | free online dictionary, thesaurus and reference materials. Released December 31, 2011. `https://www.collinsdictionary.com/`.

Atticus Harrigan and Antti Arppe. 2021. Leveraging English word embeddings for semi-automatic semantic classification in nêhiyawêwin (Plains Cree). In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 113–121, Online. Association for Computational Linguistics.

Samer Hassan and Rada Mihalcea. 2009. Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1201, Singapore. Association for Computational Linguistics.

Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to Understand Phrases by Embedding the Dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Yian Li and Hai Zhao. 2021. Pre-training universal language representation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5122–5133, Online. Association for Computational Linguistics.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

George A. Miller. 1995. WordNet: A lexical database for English. *Commun. ACM*, 38(11):39–41.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Dwaipayan Roy, Debasis Ganguly, Sumit Bhatia, Srikanta Bedathur, and Mandar Mitra. 2018. Using word embeddings for information retrieval: How

collection and term normalization choices affect performance. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 1835–1838, New York, NY, USA. Association for Computing Machinery.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Statistics Canada. 2017. *Canada [Country] and Canada [Country] (table). Census Profile.* 2016 Census. Statistics Canada Catalogue no. 98-316-X2016001. Ottawa. Released November 29, 2017. https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/index.cfm?Lang=E (accessed August 13, 2021).

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Lei Zheng, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2020. Multi-channel reverse dictionary model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:312–319.