Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1159–1166
Marseille, 20-25 June 2022

Leveraging a Bilingual Dictionary to Learn Wolastoqey Word Representations

Diego Bear, Paul Cook
Faculty of Computer Science
University of New Brunswick
{diego.bear, paul.cook}@unb.ca

Abstract

Word embeddings (Mikolov et al., 2013; Pennington et al., 2014) have been used to bolster the performance of natural language processing systems in a wide variety of tasks, including information retrieval (Roy et al., 2018) and machine translation (Qi et al., 2018). However, approaches to learning word embeddings typically require large corpora of running text to learn high quality representations. For many languages, such resources are unavailable. This is the case for Wolastoqey, also known as Malecite-Passamaquoddy, an endangered low-resource Indigenous language. As there exist no large corpora of running text for Wolastoqey, in this paper, we leverage a bilingual dictionary to learn Wolastoqey word embeddings by encoding their corresponding English definitions into vector representations using pretrained English word and sequence representation models. Specifically, we consider representations based on pretrained word2vec (Mikolov et al., 2013), RoBERTa (Liu et al., 2019), and sentence-BERT (Reimers and Gurevych, 2019) models. We evaluate these embeddings in word prediction tasks focused on part-of-speech, animacy, and transitivity; semantic clustering; and reverse dictionary search. In all evaluations we demonstrate that approaches using these embeddings outperform task-specific baselines, without requiring any language-specific training or fine-tuning.

Keywords: Word embeddings, less-resourced/endangered languages, Malecite-Passamaquoddy

1. Introduction

Pretrained word embeddings (Mikolov et al., 2013; Pennington et al., 2014) have been shown to improve the performance of natural language processing (NLP) systems for a wide variety of tasks, such as machine translation (Qi et al., 2018) and information retrieval (Roy et al., 2018). However, modern approaches to learning word embeddings typically require large amounts of running text for training in order to learn high quality embeddings. Unfortunately, for many low-resource languages, such text resources do not exist in the quantity required to obtain high quality embeddings.

Malecite-Passamaquoddy (also referred to as Maliseet-Passamaquoddy and Passamaquoddy-Maliseet) is an Eastern Algonquian language spoken in regions of what is now known as New Brunswick and Quebec, Canada, and Maine, United States. Malecite and Passamaquoddy are dialects of this language with the Malecite dialect being primarily spoken along Wolastoq (also known as the St. John River) in New Brunswick and northern Maine (Leavitt, 1996). However, many speakers of the Malecite dialect use the term *Wolastoqey* to refer to their language. This research was carried out in Wolastokuk (i.e., on or along Wolastoq, the territory in which the Malecite dialect is spoken) and the first author of this paper is wolastoqew. We therefore use the term *Wolastoqey* throughout this paper.

There are currently approximately 300 remaining first language speakers of Wolastoqey in Canada (Statistics Canada, 2017). Due to its low-resource state, developing language technologies for Wolastoqey is challenging because there are no large corpora or annotated datasets available in this language to train natural language processing systems. Despite not having large corpora or datasets available, a bilingual Wolastoqey–English dictionary, known as the Passamaquoddy-Maliseet Dictionary (Francis and Leavitt, 2008), is available. This dictionary features English definitions for Wolastoqey headwords and contains approximately $18.6k$ entries. This dictionary is also available online.[1]

As we have access to English definitions for Wolastoqey words, in our research, we propose a method to obtain word embeddings for Wolastoqey words from their English definitions by leveraging pretrained English word and sequence representation models. We do this as high quality Wolastoqey word embeddings could allow for the development of future language technologies for this language and could potentially be used to greatly increase the accessibility of existing language resources for language learners and non-speakers. We consider three types of evaluation for our Wolastoqey word representations: 1) word classification tasks focusing on predicting part-of-speech, transitivity, and animacy; 2) semantic clustering of words; and 3) reverse dictionary search. In all three types of evaluation our results demonstrate that our proposed methods for learning Wolastoqey word representations outperform task-specific baselines.

[1] Passamaquoddy-Maliseet Language Portal (http://www.pmportal.org); Language Keepers and Passamaquoddy-Maliseet Dictionary Project.

2. Related Work

There has been very little prior computational work done for Wolastoqey. To the best of our knowledge, previous research consists of two works: a preliminary finite state model of Passamaquoddy-Maliseet noun morphology (Farber, 2015), and a cross-lingual Wolastoqey–English definition modelling system that generates English definitions for Wolastoqey words (Bear and Cook, 2021). One objective of this paper is therefore to produce a system that is capable of constructing high quality Wolastoqey word embeddings that, due to the limited text resources available for this language, could not otherwise be obtained through traditional methods, and which could encourage the development of future Wolastoqey language technologies.

Previous work on Nêhiyawêwin (Plains Cree), also an Algonquian language, has demonstrated that, by leveraging a bilingual dictionary, useful vector representations can be constructed for Nêhiyawêwin words. By averaging pretrained English word2vec (Mikolov et al., 2013) embeddings corresponding to words that appear in English definitions of Nêhiyawêwin words, meaningful Nêhiyawêwin word embeddings can be obtained (Dacanay et al., 2021; Harrigan and Arppe, 2021).

Harrigan and Arppe (2021) demonstrate that, by using this approach, it is possible to obtain embeddings that can be used to semantically cluster Nêhiyawêwin words. In their work, they show that hierarchical agglomerative clustering of these embeddings cut at a specific manually-set level yields meaningful semantic clusters. Their findings indicate that this method produces particularly high quality clusters for nouns, which often require little manual effort to correct, whereas for verbs, the clustering performs notably less well.

Dacanay et al. (2021) show that this approach to deriving word representations can be applied to semantically classify Nêhiyawêwin words by mapping them to pre-constructed ontologies, namely WordNet (Miller, 1995) and RapidWords (Boerger, 2017). In the case of WordNet, a synset is represented as the average of the embeddings for its headword, and the words in its definition and synset. Nêhiyawêwin words are then mapped to the most similar synset using cosine similarity. For RapidWords, semantic domains are represented by the average of the word embeddings for words in the elicitation questions and examples. Cosine similarity is again used to map Nêhiyawêwin words to semantic domains. The results show that this approach could be used as a faster alternative to manual semantic classification, but does not achieve human-levels of semantic awareness. Dacanay et al. suggest that, in future work, BERT (Devlin et al., 2019) could be potentially used to derive a sequence representation in-place of averaging word2vec embeddings. As such, we explore using similar models in this paper.

The approach to representing Nêhiyawêwin words of Dacanay et al. (2021) and Harrigan and Arppe (2021) requires no language-specific specific information or training for Nêhiyawêwin, and as such could potentially be applied to any language with a bilingual dictionary with definitions in a high resource language. In this paper we consider whether this approach can be applied to a similar language, where both Nêhiyawêwin and Wolastoqey are Algonquian languages, which share many properties, for example the presence of polysynthesis.

3. Proposed Model

To obtain embeddings for Wolastoqey words, we experiment with encoding English definitions of Wolastoqey words in the Passamaquoddy-Maliseet Dictionary into vector representations. For this, we consider leveraging pretrained English word and sequence representation models.

3.1. Word Embeddings

The first approach we consider is similar to that of Dacanay et al. (2021). We use the average of word2vec (Mikolov et al., 2013) embeddings in an English definition to represent the corresponding Wolastoqey word. Following Dacanay et al. (2021) we use word2vec embeddings pretrained on a Google News corpus of roughly 100 billion words.[2] We obtain these embeddings using gensim 3.8.3 (Řehůřek and Sojka, 2010). This embedding model contains vector representations for roughly 3 million unique types and has a dimensionality of 300.

We preprocess definitions by removing any text that is encapsulated by brackets. This bracketed text typically provides topical information that, while potentially quite useful for a dictionary user, does not appear to contribute to the core meaning, and incorporating this text could therefore possibly be detrimental to our Wolastoqey word representations. An example of this is seen for the word *aci-tpeltomuwiw*, which is defined as '(business, building, car, painting, etc.) its ownership changes, it has new owner'. We then tokenize the definitions by splitting on non-word characters.

We represent a given Wolastoqey word as the average of the embeddings for the words in its English definition. This gives Wolastoqey word representations that have the same dimensionality as the English word embedding model, i.e., 300 dimensions. Definition words that are not in the embedding matrix are ignored. Wolastoqey words with definitions that include no words that are in the embedding matrix are removed, and are excluded from experiments using all models. In total, we consider $18k$ headwords in our experiments, as we remove 572 entries with definitions containing no words in the embedding matrix.

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

3.2. RoBERTa

We next consider using a large masked language model to obtain vector representations of Wolastoqey words from their English definitions. We do this as large-masked language models, such as BERT (Devlin et al., 2019), have been shown to achieve very strong performance on a wide range of NLP tasks, including question answering and natural language understanding. For our experiments, we use a variant of BERT known as RoBERTa (Liu et al., 2019), which is a more-robustly optimized architecture that often outperforms BERT. Specifically, we use the implementation of RoBERTa-base available in the Hugging Face Transformers 4.12.5 library (Wolf et al., 2020).

We preprocess definitions in the same manner as in Section 3 1 to remove bracketed content in definitions. Here we tokenize the input using a model-specific vocabulary, and then pass it to our RoBERTA model. This then gives outputs corresponding to each token. We use the output for the CLS token as the representation of the definition, as the CLS output is often used for sequence classification tasks and encapsulates sentence-level (here definition-level) information. This gives a vector with a dimensionality of 768, which we use as the representation of the Wolastoqey word corresponding to the definition.

3.3. Sentence-RoBERTa

Finally, we consider a specialized variant of BERT specifically pretrained for sequence representation, known as sentence-BERT (Reimers and Gurevych, 2019). We use a sentence-BERT model that makes use of the RoBERTa-base architecture, specifically the nli-roberta-base-v2 model available in the sentence transformer 2.1.0 library.[3] This model represents a pretrained checkpoint that has been trained 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 again preprocess definitions by removing bracketed text, and then apply a model-specific tokenizer. We then pass this tokenized input to the sentence-transformer model which gives us a vector representation of our input based on the mean output vectors. As this model is based on the RoBERTa-base architecture, this results in an output vector of size 768.

4. Word Classification

To evaluate how well our embeddings are able to capture syntactic properties of words, we propose using them as input to classifiers trained for a variety of word classification tasks. For this, we consider training logistic regression classifiers to predict the part-of-speech, animacy,[4] and transitivity of Wolastoqey words from their embeddings. For this, we consider using gold-standard labels from the Passamaquoddy-Maliseet Dictionary.

4.1. Experimental Setup

We consider five classification tasks focused on the following properties: 1) part-of-speech, 2) noun animacy, 3) verb animacy, 4) verb transitivity, and 5) verb type.

The $18k$ entries we use for the part-of-speech classification experiments consist of a total of 53 pronouns, 231 preverbs, 570 particles, $13.7k$ verbs and $3.3k$ nouns. For noun animacy, we ignore any nouns that occur as both animate and inanimate. This gives $1.7k$ animate, and $1.3k$ inanimate, nouns. For experiments focusing on verbs, we similarly ignore any head-word with a corresponding dictionary entry indicating that the verb can be both animate and inanimate. For verb animacy this gives $8.3k$ animate and $4.7k$ inanimate verbs. For verb transitivity this gives $5.7k$ transitive and $7.4k$ intransitive verbs. Wolastoqey verbs are categorized into four types based on the combination of their animacy and transitivity: animate intransitive, inanimate intransitive, transitive animate, and transitive inanimate. For the verb type experiments we use $5.3k$ animate intransitive, $2.1k$ inanimate intransitive, $3k$ transitive animate, and $2.7k$ transitive inanimate verbs.

The logistic regression classifiers for these experiments are implemented using scikit-learn 0.24.2 and make use of the default training parameters, with the exception of max iterations, which is set to 3000 such that all models finish converging. We evaluate our classifiers in a 10-fold cross-validation experimental setup using accuracy, as well as macro-averaged precision, recall, and F1-score. We compare against a most-frequent class baseline.

4.2. Results

Results are shown in Table 1. We observe that, for all tasks and evaluation metrics, all of our models outperform a most-frequent class baseline. This indicates that these approaches to representing Wolastoqey words capture information about these syntactic properties.

We observe that on classification tasks involving nouns (i.e., part-of-speech and noun animacy), sentence-RoBERTa (s-RoBERTa) performs best. On these tasks, RoBERTa performs worst of the three embedding approaches considered. However, RoBERTa performs best on classification tasks focused on verbs, although in many cases it only slightly outperforms our sentence-RoBERTa and word embedding models. The relatively consistent performance of sentence-RoBERTa across tasks, contrasted with the inconsistent performance of RoBERTa, could be because sentence-RoBERTa is fine-tuned to learn sequence representations, whereas RoBERTa is not.

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

[4] Wolastoqey has two grammatical genders: animate and inanimate (Leavitt, 1996).

Part of Speech				
Method	Accuracy	P	R	F1
Most Freq.	0.767	0.153	0.200	0.174
Word Emb.	0.970	**0.841**	0.705	0.743
RoBERTa	0.964	0.757	0.568	0.612
s-RoBERTa	**0.973**	0.828	**0.804**	**0.811**
Noun Animacy				
Most Freq.	0.552	0.276	0.500	0.355
Word Emb.	0.785	0.787	0.786	0.784
RoBERTa	0.731	0.733	0.724	0.725
s-RoBERTa	**0.801**	**0.800**	**0.798**	**0.798**
Verb Animacy				
Most Freq.	0.637	0.319	0.500	0.389
Word Emb.	0.962	0.959	**0.960**	0.959
RoBERTa	**0.963**	**0.961**	**0.960**	**0.960**
s-RoBERTa	0.962	0.959	0.958	0.959
Verb Transitivity				
Most Freq.	0.566	0.283	0.500	0.361
Word Emb.	0.931	0.930	0.930	0.930
RoBERTa	**0.961**	**0.960**	**0.961**	**0.961**
s-RoBERTa	0.958	0.958	0.957	0.957
Verb Type				
Most Freq.	0.406	0.101	0.250	0.144
Word Emb.	0.952	0.954	0.956	0.955
RoBERTa	**0.960**	**0.961**	**0.962**	**0.961**
s-RoBERTa	0.951	0.953	0.953	0.953

Table 1: Accuracy, precision (P), recall (R), and F1-score for each word classification task, using each embedding approach, as well as a most-frequent class baseline. The best result for each task and evaluation metric is shown in boldface.

5. Clustering

Previous work has demonstrated that similar approaches to deriving word embeddings can be used for the semi-automatic construction of semantically-organized lexicons (Harrigan and Arppe, 2021). As this is the case, here we explore using the output from our embedding models for semantic clustering of Wolastoqey words to evaluate their potential for this task.

5.1. Datasets

In-order to evaluate how well our word embeddings can be semantically clustered, we require gold-standard labels. For this, we consider using two sources, data from Wolastoqewatu,[5] an online learning platform that offers topically-organized Wolastoqey lessons, as well as data from Wolastoqey Latuwewakon,[6] a web and mobile application designed to teach Wolastoqey words and phrases through topical categories. To obtain gold-standard labels from Wolastoqewatu, we consider using the categories listed in the website's glossary. Using these categories as labels, we can extract word–category pairs. For this evaluation we remove any words that occur in multiple categories. For Wolastoqey Latuwewakon, we consider using the top-level categories from the application's categories tab. We cross reference each word we have a gold-standard label for with the content from the Passamaquoddy-Maliseet Dictionary, and keep only words that appear as headwords in the dictionary. This gives us two annotated datasets. In total, we have 1169 labelled entries from Wolastoqewatu corresponding to 20 unique classes and 79 entries with labels from Wolastoqey Latuwewakon corresponding to 6 classes.

5.2. Experimental Setup

We cluster the words in each dataset using K-means. We set the number of clusters to be the number of classes in the dataset (i.e., 20 for Wolastoqewatu and 6 for Wolastoqey Latuwewakon). We use the implementation of K-means from scikit-learn 0.24.2 with the default parameters.[7]

We evaluate the clustering using adjusted mutual information (AMI), adjusted rand index (ARI), and BCubed precision (P), recall (R), and F1-score. AMI and ARI are variants of mutual information and rand index, respectively, that are adjusted for chance. The intuition behind BCubed is to measure the quality of the clustering when a user selects one item in a cluster. High BCubed P indicates that most items in this cluster would have the same class as the selected item. High BCubed R indicates that most items with the same class as the selected item would be found in this cluster (Amigó et al., 2009). As such, BCubed might be particularly indicative of the quality of the clustering for the application of (semi-)automatically building or extending topically-focused word lists such as those in Wolastoqewatu and Wolastoqey Latuwewakon.

We compare our proposed methods to two baselines, one where all items are in one cluster, and one where all items are unique clusters.[8]

5.3. Results

Results are shown in Table 2. The results are quite consistent for both datasets. For each evaluation metric, and each dataset, the word embedding approach outperforms sentence-RoBERTa, which outperforms RoBERTa, with the exception of BCubed P for Wolastoqewatu where sentence-RoBERTa outperforms word embeddings. Furthermore, in terms of AMI, ARI, and BCubed F1, all methods outperform the baselines on both datasets, with the exception of RoBERTa on Wolastoqey Latuwewakon in terms of BCubed F1.

The majority of the words that appear in our Wolastoqewatu and Wolastoqey Latuwewakon datasets are

[5] https://wolastoqewatu.ca

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

[7] This implementation of K-means uses Euclidean distance. We also considered an implementation that uses cosine distance. The trends in the results were similar.

[8] These baselines will score 0 for both AMI and ARI.

Wolastoqewatu					
Method	AMI	ARI	BCubed P	BCubed R	BCubed F1
One cluster	0.000	0.000	0.017	**1.000**	0.034
Unique clusters	0.000	0.000	**1.000**	0.081	0.150
Word Emb.	**0.428**	**0.281**	0.339	0.410	**0.371**
RoBERTa	0.283	0.150	0.264	0.223	0.242
s-RoBERTa	0.425	0.232	0.376	0.326	0.349
Wolastoqey Latuwewakon					
One cluster	0.000	0.000	0.341	**1.000**	0.508
Unique clusters	0.000	0.000	**1.000**	0.077	0.143
Word Emb.	**0.444**	**0.354**	0.681	0.558	**0.613**
RoBERTa	0.308	0.192	0.587	0.408	0.482
s-RoBERTa	0.382	0.249	0.635	0.442	0.522

Table 2: Results for each clustering evaluation metric, for each method, on each dataset. The best result for each evaluation metric and dataset is shown in boldface.

nouns. The finding that RoBERTa performs poorly here is therefore consistent with the findings of Section 4 2, where RoBERTa performed poorly on classification tasks involving nouns. Dacanay et al. (2021) choose an approach to representing Nêhiyawêwin words using their English definitions based on word embeddings, as opposed to a transformer-based method such as BERT, because they note that the definitions they consider are often very short and non-sentential. This is also the case for the definitions in the datasets we consider in this section, which tend to be quite short. The mean token length of definitions for Wolastoqewatu and Wolastoqey Latuwewakon is four and three tokens, respectively. This could explain why the word embeddings approach outperforms sentence-RoBERTa in this evaluation. This finding is, however, inconsistent with the finding from Section 4 2 that sentence-RoBERTa outperforms word embeddings for classification tasks involving nouns. We intend to investigate this further going forward, and note that the differing dimensionalities for the models could be a factor (i.e., 300 for word embeddings and 768 for sentence-RoBERTa).

6. Reverse Dictionary

In this section we consider using Wolastoqey word representations based on English definitions to build a reverse dictionary search system. This represents a more-practical use case for our word embeddings as a high quality reverse dictionary could potentially be used to greatly increase the accessibility of language resources for Wolastoqey learners and non-speakers by making content easier to find. Our reverse dictionary is based on the principle that the English definition for a Wolastoqey word in the Passamaquoddy-Maliseet Dictionary is expected to be similar to an alternative English definition for that word.

6.1. Datasets

In order to conduct our reverse dictionary experiments we require alternative English definitions for Wolastoqey words in the Passamaquoddy-Maliseet Dictionary. Unfortunately, there exist very few other resources containing English definitions for Wolastoqey words. One such resource is the glossary contained in Wolastoqewatu. However, many of the definitions in this resource are identical to those of the Passamaquoddy Maliseet Dictionary.

Many words in the Passamaquoddy-Maliseet Dictionary have definitions consisting of a single word, especially nouns and particles. An example of this is *psuwis*, which is defined in the Passamaquoddy-Maliseet Dictionary as 'cat'. For these words which have a single-word definition, we can obtain an alternative definition using potentially any other English dictionary. We choose WordNet (Miller, 1995), primarily because of its ease of access. For each Wolastoqey word in the Passamaquoddy-Maliseet Dictionary that has a single-word definition, and that definition word also occurs as a lemma in WordNet, we use the definition for the first synset for that lemma as an alternative definition for the Wolastoqey word. Continuing with the example *psuwis*, the alternative definition is then 'feline mammal usually having thick soft fur and no ability to roar: domestic cats; wildcats'.

Nouns in Wolastoqey can be dependent (Leavitt, 1996) (i.e., inalienable). Third person forms are given for such nouns in the Passamaquoddy-Maliseet Dictionary. For example, *tus* is a dependent noun defined as 'h/ daughter', where *h/* is an abbreviation used in the Passamaquoddy-Maliseet Dictionary meaning 'his or her' in this context. Furthermore, entries for verbs are also given for third person forms. For example *unitahasin* is defined as 's/he forgets it', where *s/he* is an abbreviation which here means 'she or he'. Prior to identifying words with single-word definitions, we therefore remove all instances of *h/*, *s/he*, and *it*.

We remove any definitions containing English names in

Unrestricted Search Space								
Method	Median	Mean	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
Word Emb.	3356	0.026	0.012	0.034	0.049	0.073	0.128	0.186
RoBERTa	1615	0.017	0.004	0.018	0.043	0.064	0.108	0.141
s-RoBERTa	**107**	**0.081**	**0.027**	**0.128**	**0.183**	**0.260**	**0.397**	**0.495**
Restricted Search Space								
Random	558	0.006	0.000	0.005	0.008	0.018	0.045	0.093
Word Emb.	13	0.248	0.137	0.372	0.463	0.553	0.667	0.738
RoBerta	282	0.071	0.038	0.098	0.126	0.164	0.219	0.302
s-RoBERTa	**5**	**0.375**	**0.258**	**0.506**	**0.599**	**0.668**	**0.754**	**0.821**

Table 3: Median, mean, and accuracy@k for various thresholds, for reverse dictionary experiments using each approach to representing Wolastoqey words and a random baseline, for an unrestricted search space consisting of the entire dictionary, and a restricted search space limited to the correct answers for all queries.

the Passamaquoddy-Maliseet Dictionary using the list of English names provided in NLTK (Bird et al., 2009). This gives a total of 1091 Wolastoqey words with English definitions from the Passamaquoddy-Maliseet Dictionary and alternative English definitions from WordNet.

6.2. Experimental Setup

We represent Wolastoqey words in the Passamaquoddy-Maliseet Dictionary using their English definitions as before. We consider all headwords in this experiment except those with definitions corresponding to English names. This gives a search space of roughly 17.9k words for our reverse dictionary experiments. For each Wolastoqey word for which we have an alternative definition, we then form a representation of its alternative definition using the same preprocessing and approaches to representing definitions as for definitions in the Passamaquoddy-Maliseet Dictionary. The alternative definitions can be viewed as queries that a user might enter into a reverse dictionary system.

For each word in our dataset of Wolastoqey words with an alternative definition, we calculate the cosine similarity between the representation of this alternative definition, and the representation of each word in the Passamaquoddy-Maliseet Dictionary. We then sort the dictionary entries by their cosine similarities with the alternative definition. The Wolastoqey word corresponding to the alternative definition would then ideally be at the top of the ranking.

To evaluate our reverse dictionary we examine the rank of the Wolastoqey word corresponding to the alternative definition. Specifically, we evaluate using median rank, mean reciprocal rank (MRR), and accuracy@k, for $k = 1, 5, 10, 20, 50, 100$, in which the system is scored as correct if the word corresponding to the alternative definition is among the top-k words. We compare against a simulated random-rank baseline. In addition to using the entire dictionary as a search space, we also consider a restricted search space which only includes the 1091 Wolastoqey words which have alternative definitions.

6.3. Results

Results are shown in Table 3. For both search spaces, and all evaluation measures, we observe the same ranking of approaches: sentence-RoBERTa performs best, followed by word embeddings, then RoBERTa, and finally the simulated random baseline, with the exception that RoBERTa outperforms word embeddings in terms of median rank for the unrestricted search space. In particular, our sentence-Roberta model outperforms all other approaches by a large margin for the unrestricted search space. This finding demonstrates the potential for transformer-based approaches that have been fine-tuned for sentence representation to improve over word embedding-based approaches for representing definitions.

Despite all models outperforming the random baseline, the findings for our best model, sentence-RoBERTa, do not suggest that this could yet be used as a practical reverse dictionary search system. For example, the accuracy@100 of 0.495 of this approach when considering the unrestricted search space 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 definitions in the Passamaquoddy-Maliseet Dictionary could contribute towards making this experimental setup a particularly challenging task.

7. Conclusions

In this paper we leveraged a bilingual dictionary to represent Wolastoqey words using their English definitions. Because Wolastoqey is a low-resource language, there are no large Wolastoqey corpora available to train conventional word embedding models. Specifically, we considered approaches based on word2vec, RoBERTa, and sentence-RoBERTa. We evaluated our embeddings

based on word classification tasks focused on predicting part-of-speech, animacy, and transitivity; semantic clustering; and reverse dictionary search. In each evaluation, we found that approaches using these embeddings outperformed task-specific baselines. These findings indicate that pretrained English word and sequence representation models can be leveraged to obtain embeddings for Wolastoqey words from their English definitions that encapsulate both semantic and syntactic information.

Our results on word classification and reverse dictionary search indicate that transformer-based models fine-tuned for sequence representation can outperform approaches based on word embeddings for representing Wolastoqey words via their English definitions. In future work, we intend to further investigate transformer-based models for representing sentences, and specifically consider fine-tuning for representing dictionary definitions using monolingual English dictionaries. This could lead to improved Wolastoqey word representations while still requiring no language-specific training or fine-tuning for Wolastoqey.

8. Bibliographical References

Amigó, E., Gonzalo, J., Artiles, J., and Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486.

Bear, D. and Cook, P. (2021). Cross-lingual wolastoqey-English definition modelling. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 138–146, Held Online, September. INCOMA Ltd.

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

Boerger, B. H. (2017). Rapid word collection, dictionary production, and community well-being. In *5th International Conference on Language Documentation & Conservation*, March.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (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, September. Association for Computational Linguistics.

Dacanay, D., Harrigan, A., Wolvengrey, A., and Arppe, A. (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, June. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Farber, A. (2015). A finite-state grammar of passamaquoddy-maliseet nouns. `http://dx.doi.org/10.13140/RG.2.1.2836.6967`.

Harrigan, A. and Arppe, A. (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, June. Association for Computational Linguistics.

Leavitt, R. M. (1996). *Passamaquoddy-Maliseet*. Languages of the world. Materials, 27. LINCOM EUROPA, Munchen; Newcastle.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In Yoshua Bengio et al., editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

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

Pennington, J., Socher, R., and Manning, C. (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, October. Association for Computational Linguistics.

Qi, Y., Sachan, D., Felix, M., Padmanabhan, S., and Neubig, G. (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, June. Association for Computational Linguistics.

Řehůřek, R. and Sojka, P. (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, May. ELRA. `http://is.muni.cz/publication/884893/en`.

Reimers, N. and Gurevych, I. (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 Natu-*

ral Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China, November. Association for Computational Linguistics.

Roy, D., Ganguly, D., Bhatia, S., Bedathur, S., and Mitra, M. (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.

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).

Williams, A., Nangia, N., and Bowman, S. (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.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

9. Language Resource References

David A. Francis and Robert M. Leavitt. (2008). *A Passamaquoddy-Maliseet Dictionary*. The University of Maine Press and Goose Lane Editions.