Word-level prediction in Plains Cree: First steps

Olga Kriukova
Department of Linguistics,
University of Saskatchewan
Saskatoon, SK, Canada
olga.kriukova@usask.ca

Antti Arppe
Department of Linguistics,
University of Alberta
Edmonton, AB, Canada
arppe@ualberta.ca

Abstract

Plains Cree (nêhiyawêwin) is a morphologically complex and predominantly prefixing language. The combinatory potential of inflectional and derivational/lexical prefixes and verb stems in Plains Cree makes it challenging for traditional auto-completion (or word suggestion) approaches to handle. The lack of a large corpus of Plains Cree also complicates the situation. This study attempts to investigate how well a BiLSTM model trained on a small Cree corpus can handle a word suggestion task. Moreover, this study evaluates whether the use of semantically and morphosyntactically refined Word2Vec embeddings can improve the overall accuracy and quality of BiLSTM suggestions. The results show that some of the models trained with the refined vectors provide semantically and morphosyntactically better suggestions. They are also more accurate in predictions of content words. The model trained with the non-refined vectors, in contrast, was better at predicting conjunctions, particles, and other non-inflecting words. The models trained with different refined vector combinations provide the expected next word among top-10 predictions in 36.32 to 37.34% of cases (depending on the model).

1 Introduction

Auto-complete systems and predictive text input have become integral components of our daily interactions with our devices and digital platforms. These applications heavily rely on robust language models capable of accurately predicting the next word in a given sequence of text. While substantial progress has been made in developing efficient language models for major languages, the challenges persist for low-resource languages where scarcity of training data poses a significant obstacle. This challenge is especially found for Indigenous languages that are often also morphologically rich.

With advances in the NLP and machine learning fields, small training datasets have become less of a problem; however, the handling of the morphological complexity still presents a challenge. Lane and Bird (2020) approached this problem with the development of an interactive word-completion system for Kunwingku (an Indigenous language spoken in Northern Australia) based on a finite state recognizer which included most morphology for some 500 verbs. Their tool suggests a completion up to the next morpheme boundary and helps to avoid the so-called "combinatorial explosion of possible words" typical for the prefixing polysynthetic languages.

Lane et al. (2022) further successfully extend this method to Plains Cree, with a full-fledged model including all parts of speech, covering most inflectional morphology for the inflecting verbs and nouns, and based on a lexicon of well over 20k lexemes. The tool is based on a finite state morphosyntactic analyzer of Plains Cree (nêhiyawêwin, an Indigenous language spoken mainly in on the Western Canadian Plains) (Snoek et al., 2014; Harrigan et al., 2017). It uses corpus-based information about Cree prefixes to predict the most probable and common next morpheme in a word (based on a small corpus of some 150k Cree words). While the results were perceived as surprisingly good, given the small size of the corpus, there remained yet quite many valid optional completions, since their tool did not make use of preceding lexical or morphosyntactic context. Similarly, with Lane and Bird (2020), Plains Cree interactive word completion could be used by non-fluent Cree speakers and learners who may struggle to build word forms. Nevertheless, as the successful use of the model requires a broad knowledge of the language and its word formation, in order to be able to choose the completion appropriate to the context, they considered that the compilation model might be most useful for fluent speakers. Additionally, the model is helpful to fluent speakers who have difficulties with diacritics for vowel length and other aspects

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 15–23

of spelling, in support of which they also included a spelling correction component.

The present research draws inspiration from these pioneering works of Lane et al. (2022) and Lane and Bird (2020), and seeks to continue the experiments in the field of word completion for Plains Cree. The present study, however, aims to investigate the feasibility of a complete word prediction and seeks to provide fluent Cree speakers with morphosyntactically and contextually appropriate, if not accurate, word suggestions which can potentially speed up the typing process.

To achieve this, we train a Bidirectional Long Short-Term Memory (BiLSTM) model to predict the next word in a sequence. LSTMs, a type of recurrent neural network (RNN), have demonstrated remarkable success in capturing contextual dependencies in sequential data, making them a compelling choice for natural language generation tasks (Hochreiter and Schmidhuber, 1997; Sundermeyer et al., 2012). Moreover, the fact that several studies working with low-resource agglutinative and polysynthetic languages used LSTMs for word prediction task (Kosyak and Tyers, 2022), makes it a compelling choice for the Plains Cree case.

To improve the model performance, we also train Word2Vec embeddings (Mikolov et al., 2013) for words in the Plains Cree corpus (see Section 2.2). Additionally, in this study, we explore the effect of vector augmentation—based on the words' morphosyntactic analyses, WordNet semantic classes (if applicable) (Miller, 1995), and lemmas—on the overall model accuracy and quality of word suggestions.

The paper is structured as follows. The data used in this paper and data preprocessing are described in Section 2. The Word2Vec vectors training and refinement, and BiLSTM model training are described in Section 3. Section 4 presents the results, and Section 5 presents their discussion. Possible directions for further research are examined in Section 6.

2 Data

2.1 Plains Cree

Plains Cree (endonimically – nêhiyawêwin, ISO 639-3: crk) is an Algonquian language spoken in Alberta, Saskatchewan, and Northwest Territories in Canada, and in the northern part of Montana in the US. This is the most widely spoken dialect of Cree. Cree is an agglutinative and polysynthetic language of predominantly prefixing nature. Although Cree is among the most spoken Indigenous languages of Canada, only a small corpus of Plains Cree is available currently.

2.2 Plains Cree corpus

The training data for this study comes from a combination of the Ahenakew-Wolfart Corpus (Arppe et al., 2020), the Bloomfield Corpus (Schmirler, 2023), and the Corpus of Miscellaneous Plains Cree Texts (*misi-mîkiwâhp pêsêkinosa ohci*) (Dacanay and Arppe, 2024), which has been morphosyntactically analyzed and lemmatized with the finite-state model mentioned before (Snoek et al., 2014; Harrigan et al., 2017), morphosyntactically disambiguated with a CG parser (Schmirler et al., 2018; Schmirler, 2023), and then annotated for WordNet semantic class (for nouns and verbs, when available). The WordNet classes are based on the classification by Dacanay (2022) of over 20k Cree entries in the lexical database underlying the *Cree Words/nêhiyawêwin: itwêwina*, a bilingual English-to-Cree dictionary by Wolvengrey (2001). All this information about each word was organized in a .tsv file, where each row included word form, analysis, and WordNet class as shown below:

awâsisak awâsis+N+A+Pl (n) child#1

2.3 Preprocessing

Before the corpus could be used for training it required significant preprocessing. First, the standard corpus normalization steps were made: 1) punctuation signs were removed, 2) words were converted to lowercase, and 3) Arabic and Roman numerals were removed.

Secondly, some special notes from the corpus were taken out. They include speakers' initials, new segment markers (e.g., 'Part'), web URLs for the texts taken from the Internet pages, and transcribers' notes (e.g., 'laughs', 'gesture').

Thirdly, English pieces were removed from the corpus along with the personal names. Lastly, the word 'and', connecting multiple WordNet classes, were removed, while all the spaces were replaced with underscore signs, as exemplified below:

(n) bannock#1 and (n) bread#1 and (n) flour#1

-> (n)_bannock#1_(n)_bread#1_(n)_flour#1

For words that lack WordNet class (e.g., conjunctions, pronouns) the UNK (unknown) code was added.

Lastly, the morphosyntactic analyses were also preprocessed. Originally, each analysis contained

	W/o preverbs	With preverbs
Word tokens	224,440	281,269
Word types	50,313	40,404
Lemmas	28,200	28,250
WordNet types	5,540	5,545
Analysis types	4,595	4,637

Table 1: Training dataset features.

a word lemma, for instance:

kâwiy+N+A+Px1Sg+Sg *or*

PV/ki+***itêw***+V+TA+Ind+X+3SgO

Lemmas were extracted from the morphosyntactic analyses, to provide an additional source of information about each word, resulting in the following representations:

+N+A+Px1Sg+Sg *or*

PV/ki+V+TA+Ind+X+3SgO

All the manipulations with the corpus were done with regular expressions.

Next, we separated verbs from preverbs for easier processing. In Plains Cree, preverbs are a broad category that includes both grammatical and derivational/lexical morphemes. As their name suggests, they appear before the verb stem. Plains Cree verbs can have multiple preverbs attached. Preverbs are usually separated from each other and the verb stems by hyphens (Okimâsis, 2004, 17). For instance, *nikakwê-nêhiyawân* 'I try to speak Cree', where the preverb *kakwê-* means 'try to, attempt to'. The number of combinations that preverbs can form is enormous, as shown by the Lane et al. (2022). Therefore, we decided to separate preverbs from their stems and treat them as separate entries in the training dataset for the purposes of this study. By doing so, we expect to improve our model prediction accuracy, because it will be able to learn preverbs combinations and their relations with different verb stems.

2.4 Training dataset

After all the aforementioned preprocessing steps, the dataset presented in Table 1 was obtained. The left column shows the size of the corpus before preverb separation and the right column - after separation.

3 Language modelling

3.1 Word2Vec pre-training

To improve the performance of the LSTM model, we decided to pre-train Word2Vec vectors using the CBOW approach. We experimented with different window sizes and settled with window size 5, because it was giving the best results. Considering the amount of information about each word available in the dataset, we decided to make the most of it during the Word2Vec pre-training. In order to do so, separate vectors were trained on the sequences of words[1], their WordNet semantic classes, their lemmas, and their morphosyntactic analyses, giving us four sets of vectors. The average of four vectors was calculated, and the original vectors for words were updated with the refined ones. Thus, the final word vectors are based not only on the neighbouring words but also on the semantic classes, lemmas, and morphosyntactic features of these neighbouring words.

Some adjustments had been made, however, to address the case of the words without WordNet class (such as non-inflecting particles) and/or morphosyntactic analysis. The vector refinement was organized in such a way that vectors for the 'Unknown' WordNet class or analysis were not included in the vectors' averaging. So, for example, the refined vector for the word *mêskanaw* (mêskanaw; N+I+Sg; (n)_road#1_(n)_trail#2) was the average of the vectors of *mêskanaw* and its lemma, analysis, and semantic class. However, for the particle *iyikohk* ('to such a degree; to such an extent'), the WordNet class is unavailable, so its averaged vector is based on the word, analysis, and lemma vectors only.

3.2 Model training

We trained a BiLSTM language model using the pre-defined word vectors as embeddings. As was mentioned earlier, the Bidirectional Long Short-Term Memory model was chosen for the purposes of this study. The LSTM model was chosen for our experiments because it was previously successfully used for predictive text tasks for polysynthetic low-resource languages (Kosyak and Tyers, 2022). At the beginning of the study, both unidirectional and bidirectional models were tested, and BiLSTM

[1]For convenience, we will be using the term 'word' to refer to all the Cree tokens in the training dataset, which also include preverbs.

showed better results. Thus, we decided to proceed with BiLSTM.

The data for language modelling was split into modelling and testing subsets at a ratio of 90% to 10%. The proportionately larger training set is due to the relatively small amount of data. The models' evaluation was based on the 3-fold cross-validation results.

The Word2Vec training parameters are provided in Appendix A. The model training hyperparameters are provided in Appendix B. All the BiLSTM models were trained on the university's high-performance computing cluster. The average training time for each model was 2-3 hours.

Vectors	Correct prediction		
	Top 1	Top 5	Top 10
Non-refined	15.06	31.15	**38.55**
Word+Analysis	14.19	29.92	**37.34**
Word+WordNet	13.92	29.59	37
Word+Analysis +Lemma	13.8	29.57	37.14
Word+Analysis +WordNet	13.67	29.17	36.54
Word+WordNet +Analysis+Lemma	13.46	28.86	36.32

Table 2: Prediction accuracy for the models trained with different Word2Vec sets (mean scores of 3-fold cross-validation).

4 Results

4.1 Refined BiLSTM model

To evaluate our model we looked at its top 10 predictions. The number 10 was chosen because it covers the number of suggestions provided by all commonly used tools, such as a smartphone keyboard (3-9 suggestions), search engine (9 suggestions), and text editors (5-9 suggestions). The resulting model predicted the correct word in the top 10 predictions in 36.3% of cases. Mostly, the model handled the prediction of nouns, adverbs, and preverbs better than of verbs. In comparison to the model trained with non-refined vectors, the overall quality of predictions slightly improved, and suggestions became more contextually and grammatically suitable (see Section 4.2 for comparison).

In cases when the model could not provide the correct completion among the first 10 predictions, the first letter of a word was provided to the model, simulating the beginning of user input. As a result, 41% of previously non-predicted words were eventually suggested as a possible completion in the top 10 predictions. The majority of the words predicted with the first letter input were verbs.

To sum up, the model predicts the next word by preceding context in 36.3% of cases and predicts the next word by context and the first letter in 28% of cases. However, as will be shown in the next section, this version of the model was not the most accurate.

4.2 Model performance with different Word2Vecs

To evaluate how the refined word vectors affected the overall model performance, we conducted several experiments with different combinations of refined vectors. It was mainly done to scrutinize how information about semantic, morphosyntactic, and lemma sequences contributes to the accuracy of the BiLSTM model. Table 2 shows the results of these experiments with a percentage of correct predictions in the top one, top 5, and top 10 predictions. It should be noted that these numbers do not evaluate the semantic and morphosyntactic appropriateness of the suggestions.

As can be seen from Table 2, the best accuracy of predictions was shown by the model that was trained with non-refined word vectors. However, the results of the models trained with different sets of refined vectors are not dramatically different as well. Nevertheless, it is interesting to compare the performance of the refined vectors' models. First, the averaging with lemma vectors does not seem to provide better prediction results. In both cases when they were used the overall accuracy dropped in comparison to the same vectors' combinations without lemma. The morphosyntactic analysis, on the contrary, seems to provide valuable information about the word's neighbours. The Word+Analysis model provides the best results (37.34%) among the models with refined vectors. The next best result is shown by the Word+Analysis+Lemma model (37.14%). The model trained with the full Word2Vec set shows the lowest accuracy results.

To better understand how the suggestions changed with the refined vectors, we compared the models' results. We began with a comparison of the predictions produced by the BiLSTM trained with the Word+WordNet+Analysis+Lemma refined vectors (hereafter full BiLSTM) and the one trained with the simple word vectors (hereafter BiLSTM).

This comparison revealed that the full BiLSTM had better accuracy in predicting nouns and verbs, in contrast to other parts of speech that do not have WordNet class. The overall top-10 suggestions became more semantically and grammatically suitable to the context than those predicted by BiLSTM. In some cases, predictions of full BiLSTM were not equal to the originally occurring word, but they all were in the correct morphosyntactic form. For example, in one case, the word *okimâwa* 'another chief' was expected, and the full BLSTM had it as a top-10 prediction, but it also offered *iskwêwa* 'another woman', *oskinîkiwa* 'another young man', *nâpêwa* 'another man', *mostoswa* 'another cow'. Most of them (except the 'cow') are semantically close to the expected word and denote humans, and all of them are in the expected obviative form. The top-10 BiLSTM's predictions for this case there were also some relevant suggestions in the correct morphosyntactic form (e.g., *oskinîkiwa*, *iskwêwa*, *mostoswa*); however, it had more semantically unsuitable suggestions than full BiLSTM like *mostoswa* 'another cow', *wâkayôsa* 'another black bear', *êskana* 'another antler', *misatimwa* 'another horse', *ôhi* 'this one'. Another example like this is illustrated in Table 3. Both models predicted the correct word in the top 3 suggestions, but the overall prediction quality is better in the full BiLSTM case. Interestingly, both Word+Analysis and Word+WordNet models predicted this word as top-1 and did not consider the preverb 'kâ-' in top-5 suggestions. Word+Analysis model predictions were all in the correct morphosyntactic form. Expectedly, Word+WordNet predictions were better semantically sorted, but not all of them were in the obviative form.

Another example (see Table 4), represents how some of the suggestions of the full BiLSTM, although a bit 'ambitious' do not sound completely absurd as well. All of them fit the overall structure of the sentence, and some words fit the context nicely (e.g., gift, decision). The 'ôma' suggestion after the 'ôma' in the preceding context is, most likely, a result of the word repetitions natural for spoken language presented in the corpus. The Word+Analysis BiLSTM also predicted the expected word, but other suggestions were less satisfactory. The simple BiLSTM did not predict the next word in this sequence.

In the case of verb predictions, we can observe a more or less similar situation. Full BiLSTM provides better suggestions than simple BiLSTM. For instance, in the case of the example provided in Table 5 the full BiLSTM correctly predicts the next verb *âtotamân* 'S/he will tell' in the top 3 predictions. The regular BiLSTM, in contrast, could not provide the expected verb. Interestingly, all the other refined models predicted the correct verb, with Word+Analysis+WordNet and Word+Analysis models providing the best suggestions. The Word+Analysis model's suggestions are also provided in Table 5.

Moreover, the separation of the preverbs from the verb stems allowed all the models to suggest out-of-vocabulary preverb+verb combinations. However, the Word+Analysis refined model offered the best preverbs and verb suggestions. When the first preverb was provided, for example, 'ê-', this model suggested possible next preverbs (e.g., kî-), as well as possible verbs, to follow it. It is also important to note that some of the predictions are only possible, if they are not prefixed, as they incorporate initial change, e.g. êtwêhk, êtwêt, and êtât. Thus, further work is needed to address these cases.

Finally, we observed that the refined models were less effective in the prediction of particles. The simple BiLSTM was on average 10% more successful in predicting them (e.g., *awa* or *ôma* 'this', *êkosi* 'so, thus'). Moreover, refined models often failed to predict low-frequency words that, in addition, did not have a WordNet class and morphosyntactic analysis assigned in our training dataset.

5 Discussion and Conclusions

Although testing of the models shows that the overall accuracy is higher for the simple BiLSTM, we argue that these results need further analysis and discussion before we can come to the final conclusion about vector refinement efficiency for Plains Cree word prediction. In order to quantify the results of our qualitative observations, we did two additional tests on the out-of-fold prediction results. First, we analyzed how semantically close were the predictions to the expected word. Second, we measured the morphosyntactic similarity of the predictions and the expected words.

For the first test, we measured a Wu-Palmer Similarity between the WordNet classes of the predictions and the target words with the NLTK WordNet package. The Wu-Palmer similarity value repre-

Input: *...cêskwa! itwêw awa sihkihp. êkotê isi kapâw. miyosiyiwa ôhi _*		
Eng: Wait! S/he says this is a waterhen. Towards there, s/he goes ashore. Someone is beautiful, this is_'		
No.	**Full BiLSTM predictions**	**BiLSTM predictions**
1	*ka-* Preverb	*kâ-* Preverb
2	***oskinîkiskwêwa*** 'another young woman'	*iskwêwa* 'another woman'
3	*iskwêwa* 'another woman'	***oskinîkiskwêwa*** 'another young woman'
4	*oskinîkiwa* 'another young man'	*ê-* Preverb
5	*wâkayôsa* 'another black bear'	*oskinîkiwa* 'another young man'
6	*nâpêwa* 'another man'	*nâpêwa* 'another man'
7	*kisêyiniwa* 'another old man'	*another wâkayôsa* 'another black bear'
8	*nâpêsisa* 'another boy'	*mostoswa* 'another cow'
9	*okimâwa* 'another chief'	*kisêyiniwa* 'another old man'
10	*nôtokêsiwa* 'another old woman'	*okimâwa* 'another chief'

Table 3: Predictions comparison 1

Input: *...pîhci ôma owiyasiwêwin piko ta-kawotinikêhk ôma _*		
Eng: 'By law, s/he must take back this _'		
No.	**Full BiLSTM predictions**	**BiLSTM predictions**
1	*miyikosiwin* 'gift'	*ôma* 'this'
2	CRTC	*ka-* Preverb
3	***askiy*** 'land'	*owiyasiwêwin* 'law, decision'
4	*wîhtamâkêwin* 'statement, announcement'	*nêhiyaw* 'Cree person'
5	*ôma* 'this'	*mâmiskôcikâtêwin* 'discussion'
6	*pîkiskwêwina* 'words'	*kistêyihtcikâtêwin* 'importance; principle'
7	*tahtoskânêsiwak* 'United Nations'	***askiy*** 'land'
8	*owiyasiwêwin* 'law, decision'	*wîhtamakêwin* 'statement'
9	*mâmawâyâwinihk* 'community, group'	*wiyastêwin* 'context, foundation'
10	*wiyastêwin* 'structure, arrangement, format'	*miyo-âyâwin* 'prosperity, good health'

Table 4: Predictions comparison 2

Input: *...âcimowin ôma k-ôh-nitotamâkawiyân k- _*		
Eng: 'This story, you (pl.) have not told me _'		
No.	**Full BiLSTM predictions**	**Word+Analysis BiLSTM predictions**
1	*ôh-* Preverb	*ôh-* Preverb
2	*ayâyân* 'I will say'	*âti-* Preverb
3	***âtotamân*** 's/he will tell'	***âtotamân*** 's/he will tell'
4	*êsiyîhkâtêk* 'it will be called'	*êtwêt* 's/he says so'
5	*âtotamân* 'you will tell us about it'	*êtwêhk* 'people say'
6	*êtwêhk* 'people say'	*êtât* 'you (sg) say thus to him'
7	*êtwêt* 's/he will say'	*êsiyîhkâtêk* 'it will be called'
8	*ês-âsotamawiyâhk* 'you (sg) promise to us'	*êtwêyân* 'I will say'
9	*ây-* Preverb	*âyâyâhk* 'for us to be there'
10	*êtât* 'you will say to him/her'	*êsiyîhkâsot* 's/he is called so'

Table 5: Predictions comparison 3

Model	Wu-Palmer similarity	MorphSyn similarity
Non-refined	37.9	42.17
Word+Analysis	**38.22**	**42.5**
Word+WordNet	37.86	41.86
Word+Analysis +Lemma	37.93	42.21
Word+Analysis +WordNet	37.48	42.1
Word+WordNet +Analysis+Lemma	37.78	42

Table 6: Average Wu-Palmer similarity and morphosyntactic (MorphSyn) similarity of the out-of-fold predictions to the actual labels

sents the distance between two synsets within the WordNet semantic hierarchy tree. It ranges from 0 to 1; the higher the value the more semantically similar two words are. The second column of Table 6 shows the average semantic similarity of top-10 predictions made by the models to the corresponding target words. The similarity was counted for/with applicable words only, i.e. words marked for the WordNet class.

For the second test, we calculated the Jaccard coefficient index for the morphosyntactic analyses of target words and predicted words. This comparison intended to show how well the suggested words were able to fit the morphosyntactic structure of the sentence. The third column of Table 6 demonstrates the average Jaccard similarity index of all the predictions and target words' pairs. Similarly, with the first test, the similarity index was calculated only for the words with the morphosyntactic analysis in the dataset.

The results of the tests showed that Word+Analysis and Word+Analysis+Lemma refined models provided contextually and grammatically better suggestions in comparison to other models. An average suggestion of the Word+Analysis model was 42.5% grammatically and 38.22% semantically similar to a target word. However, the difference is too small to claim with certainty that the refined vectors significantly improved the quality of predictions.

Although vector refinement does not provide a substantial prediction improvement, the main experiment and the additional tests indicate that the morphosyntactic information about words contributed the most to the refined models' accuracy and quality of suggestions. Probably, this result is due to the lower number of words lacking morphosyntactic analysis (in comparison to words lacking WordNet class). The lemma information does not seem to contribute to the overall accuracy of prediction. Nevertheless, it seems to improve the overall quality of suggestions. The information about neighbouring WordNet classes did not improve the accuracy or quality of word prediction. However, it is most likely related to the high number of words in the dataset that were not assigned a WordNet class yet or do not have a WordNet class.

The analysis of the models trained with refined vectors and the regular model revealed a disadvantage of embeddings refinement in the present settings. Although refined vectors contributed to the slight prediction quality improvement for the words that had all the extra information like morphosyntactic analysis, WordNet class, and lemma, they did not provide an adequate representation for the words lacking some or all of this extra information. The vector refinement function did not update its vectors based on the additional information. Consequently, after refinement, they could appear further from the words they originally co-occurred with, because their 'neighbours' vectors were updated. This highlights the necessity for a better refinement approach in further studies. In this study we used simple vector averaging, but in the follow-up studies the more sophisticated approaches like those proposed by Faruqui et al. (2015) and Mrkšić et al. (2017) should be explored.

To interpret the results of this study, it is also important to keep in mind that the model was tested on a small chunk of the corpus. Our corpus, in general, has a large portion that comes from transcribed Plains Cree narratives and fiction stories. Transcribed narratives often have more filler words, while fictional stories often have rare literary words. Both significantly differ from the writing we use on a day-to-day basis (texting, search queries, etc.). In some cases, the preceding context may have many words without semantic class and that makes predictions of the following words very tricky. Thus, we are sure that under the present circumstances, the experiment with refined BiLSTM models training yielded promising results. In future, we want to experiment with more standardized texts for training and testing and explore the possibility of excluding filler words for the Word2Vec training. More-

over, we believe that the predictions' quality can be also improved by reducing the number of non-analyzed words with unknown WordNet classes in our corpus. Further improvements in the training dataset can allow the model to learn more about contextual neighbours of each word, WordNet class and morphosyntactic analysis.

To conclude, this research lays the groundwork for a future predictive text model for Plains Cree. It shows that full-word prediction is not impossible for Plain Cree, and with certain improvements and modifications, can reach higher accuracy levels. This study also explores the use of augmented word embeddings in data scarcity cases; however, the efficiency of this method requires further analysis with a fuller dataset. Potentially, use of the model in tandem with other rule-based tools and resources developed for Plains Cree, such as morphosyntactic analyzer (Snoek et al., 2014; Harrigan et al., 2017), constraint grammar parser (Schmirler, 2023), or weighted Plains Cree morpheme combinations (Lane et al., 2022), can lead to more accurate results. Naturally, significant improvements are required before speakers and learners can use this tool.

6 Future work

Naturally, this study is only the beginning of the journey to the full-scale tool for the predictive text for Plains Cree. Hence, there are several directions for future research and experiments that we plan to pursue next to address the gaps in the present study.

First, we would like to try using fastText embeddings (Bojanowski et al., 2017) to capture regularities on the sub-word level of Plains Cree beyond preverbs. FastText embeddings were already successfully used for Mi'kmaq (Boudreau et al., 2020), another Algonquian language, and provided substantial improvements to the Mi'kmaq word prediction model. It would be interesting to compare the results of the word prediction model trained with refined word vectors used in this study, and the one trained with fastText embeddings. Hypothetically, the fastText-based language model should handle Plains Cree verbs better, because it will be able to capture other aspects of rich Plains Cree morphology (like verbal suffixes), by learning it on a sub-word level.

Secondly, we would like to implement partial input suggestions beyond preverbs for long morphologically complex words (e.g., verbs with 3+ preverbs). This approach will require a different methodology for dataset preparation and prediction assessment. Moreover, the keystroke saving tests will be required to explore the efficiency of partial input suggestions for keyboard users.

Limitations

Since this study is experimental, many problems have not been addressed here yet. First of all, our model does not have mechanisms to work with OOV (out-of-vocabulary) words. It only knows the words and prefixes it encountered in the corpus. This significantly limits the model at this stage, however, we plan to address this issue during the next development phase. Secondly, the model has difficulties in predicting longer and more morphologically complex words. As mentioned above, we plan to fix this by implementing partial input predictions. Thirdly, the model does not yet have a spelling relaxation function that would allow users to type without diacritics and still get predictions.

Ethics Statement

The tools described in this manuscript have been developed in order to support the explicit objectives of the language communities in question, to support their language instruction, maintenance, and revitalization activities.

Acknowledgements

We are grateful to the Plains Cree (nêhiyawak) communities, in particular those associated with the Maskwacîs Education Schools Commission (MESC) in Maskwacîs, Alberta, Canada, for the opportunity to work with their language. This research was supported in part by a Partnership Grant (#895-2019-1012) from the Social Sciences and Humanities Research Council (SSHRC) of Canada.

We also want to thank our anonymous reviewers for their valuable feedback that helped strengthen the study's quality.

References

Antti Arppe, Katherine Schmirler, Atticus G Harrigan, and Arok Wolvengrey. 2020. A morphosyntactically tagged corpus for plains cree. In *49th Algonquian Conference (PAC49)*, volume 49, pages 1–16.

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.

Jeremie Boudreau, Akankshya Patra, Ashima Suvarna, and Paul Cook. 2020. Evaluating the impact of subword information and cross-lingual word embeddings on Mi'kmaq language modelling. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2736–2745.

Daniel Dacanay. 2022. *A Comparative Analysis of Manual and Vector Semantic Organisation using a Bilingual Dictionary of Plains Cree*. Bachelor's Thesis, University of Alberta, Edmonton, Canada.

Daniel Dacanay and Antti Arppe. 2024. misi-mîkiwâhp pêsêkinosa ohci – A corpus of miscellaneous Plains Cree texts. In *Papers of the 55th Algonquian Conference (PAC55)*.

Manaal Faruqui, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Edward Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies*, pages 1606–1615.

Atticus G Harrigan, Katherine Schmirler, Antti Arppe, Lene Antonsen, Trond Trosterud, and Arok Wolvengrey. 2017. Learning from the computational modelling of plains cree verbs. *Morphology*, 27:565–598.

Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735 – 1780.

Sergey Kosyak and Francis M. Tyers. 2022. Predictive text for agglutinative and polysynthetic languages. In *Proceedings of the First Workshop on Field Linguistics*, pages 77–85.

William Lane and Steven Bird. 2020. Interactive word completion for morphologically complex languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4600–4611.

William Lane, Atticus Harrigan, and Antti Arppe. 2022. Interactive word completion for Plains Cree. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 1:3284–3294.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

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

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics Transactions of the Association for Computational Linguistics*, 5:309–324.

Jean L. Okimâsis. 2004. *Cree: Language of the Plains/nêhiyawêwin: paskwâwi-pîkiskwêwin.* University of Regina Press.

Katherine Schmirler. 2023. *Syntactic Features and Text Types in 20th Century Plains Cree: A Constraint Grammar Approach*. PhD Dissertation, University of Alberta, Edmonton, Canada.

Katherine Schmirler, Antti Arppe, Trond Trosterud, and Lene Antonsen. 2018. Building a Constraint Grammar Parser for Plains Cree Verbs and Arguments. In *Proceedings of the 11th Language Resources and Evaluation Conference*, pages 2981–2988.

Conor Snoek, Dorothy Thunder, Kaidi Lõo, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2014. Modeling the Noun Morphology of Plains Cree. In *ComputEL: Workshop on the use of computational methods in the study of endangered languages*.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. LSTM neural networks for language modeling. *Interspeech*.

Arok Wolvengrey. 2001. *nêhiyawêwin: itwêwina / Cree: Words (Bilingual edition)*. University of Regina Press.

A Word2Vec hyperparameters

The following hyperparameters were used to train Word2Vec embeddings.

Parameter	Value
vector size	300
window	5
min count	1
workers	4

Table 7: Word2Vec hyperparameters

B BiLSTM hyperparameters

The following BiLSTM hyperparameters provided the best training results.

Parameter	Value
BiLSTM layers	3
embedding dim	300
layers dropout	0.3
sequence length	21
optimizer	Adam
learning rate	0.001

Table 8: BiLSTM hyperparameters