Morphological Segmentation for Seneca

Zoey Liu
Boston College
ying.liu.5@bc.edu

Robbie Jimerson
Rochester Institute of Technology
rcj2772@rit.edu

Emily Prud'hommeaux
Boston College
prudhome@bc.edu

Abstract

This study takes up the task of low-resource morphological segmentation for Seneca, a critically endangered and morphologically complex Native American language primarily spoken in what is now New York State and Ontario. The labeled data in our experiments comes from two sources: one digitized from a publicly available grammar book and the other collected from informal sources. We treat these two sources as distinct domains and investigate different evaluation designs for model selection. The first design abides by standard practices and evaluates models with the in-domain *development set*, while the second one carries out evaluation using a *development domain*, or the out-of-domain development set. Across a series of monolingual and cross-linguistic training settings, our results demonstrate the utility of neural encoder-decoder architecture when coupled with multi-task learning.

1 Introduction

A member of the Hodinöhšöni (Iroquoian) language family in North America, the Seneca language is spoken mainly in three reservations located in Western New York: Allegany, Cattaraugus and Tonawanda. Seneca is considered acutely endangered and is currently estimated to have fewer than 50 first-language speakers left, most of whom are elders. Motivated by the Seneca community's language reclamation and revitalization program, a few hundred children and adults are actively learning and speaking Seneca as a second language.

To further facilitate the documentation process of Seneca, recent years have witnessed the scholarly bridge between the language community and academic research, taking advantage of rapidly evolving technologies in natural language processing (NLP) (Neubig et al., 2020; Jimerson and Prud'hommeaux, 2018). In particular, ongoing work has mainly been devoted to developing automatic speech recognition (ASR) systems for Seneca (Thai et al., 2020, 2019). Their findings demonstrated that when combined with synthetic data augmentation and machine learning techniques, robust acoustic models could be built even with a very limited amount of recorded naturalistic speech. More importantly, the research output was incorporated into the Seneca people's documentation endeavors, illustrating the potential of collaborations between language communities and academic researchers.

The current study contributes to this line of research with the same ethical considerations (Meek, 2012). Specifically, we focus on morphological segmentation for Seneca, an area that has not yet been investigated thus far. Given a Seneca word, the task of morphological segmentation is to decompose it into individual morphemes (e.g., *hasgatgwë's* → *ha* + *sgatgwë'* + *s*).

With a series of in-domain, cross-domain and cross-linguistic experiments, the goal of our work is to build effective segmentation models that can support the community's ongoing language reclamation and revitalization efforts. Particularly for morphologically rich languages, it has been shown that morphological segmentation is a useful component in certain NLP tasks such as machine translation (Clifton and Sarkar, 2011), dependency parsing (Seeker and Çetinoğlu, 2015), keyword spotting (Narasimhan et al., 2014), and automatic speech recognition (ASR) (Afify et al., 2006). Given that Seneca is a highly polysynthetic language (see Section 2), good morphological segmentation models show promise for the development of other computational systems such as ASR, which would facilitate the documentation process of the language itself.

Another motivation for our experiments lies in the fact that previous research on morphological segmentation has mostly concentrated on

Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 90–101

Indo-European languages in high-resource settings (Goldsmith, 2001; Goldwater et al., 2009; Cotterell et al., 2016b), sometimes relying on external large-scale corpora in order to derive morpheme or lexical frequency information (Cotterell et al., 2015; Ruokolainen et al., 2014; Lindén et al., 2009). By contrast, work on morphological segmentation of augmented low-resource settings or truly under-resourced languages is lacking in general (Kann et al., 2016). Hence demonstrations of what model architecture and training settings could be beneficial with data sets of very small size would be informative to other researchers whose work shares similar goals and ethical considerations as ours.

2 Data Statements

Following recently advocated scientific practices (Bender and Friedman, 2018; Gebru et al., 2018), we would like to first introduce the data of the indigenous languages to be explored.

The protagonist in our experiments is Seneca, the data of which came from three sources: the book *The Seneca Verb: Labeling the Ancient Voice* by Bardeau (2007) [1], informal transcriptions provided by members from the community, and a recently digitized Bible translated into Seneca. The grammar book provides morphological segmentation for only verbs and the morpheme boundaries were based on rules defined by grammarians. By contrast, the informal sources contain labeled segmentation for a mix of verbs and nouns conducted by community speakers. The Bible offers only unlabeled data.

One of the most distinct features of Seneca morphology is that it is highly polysynthetic. This means that a single word can consist of multiple morphemes and may contain more than one stem; and this single word is able to express the meaning of a whole phrase or even sentences at times (Aikhenvald et al., 2007; Greenberg, 1960). As a demonstration, consider the following example (the indicated morphological characteristics here abide by the annotation standards of Sylak-Glassman (2016)). Breaking the Seneca word into individual morphemes, *ye:nö* is the stem which has the verbal meaning of *grab* in present tense; the prefix *ke* denotes that *ye:nö* is a transitive action, with *I* being the subject and *her/them* being the object; the single apostrophe ' at the end marks the stative state.

(1) keyenö'

ke yenö '
I+her/them grab STAT

I've grabbed her/them.

A large number of words in Seneca have agglutinative morphological features, meaning when multiple morphemes are combined during word formation, their original forms remain unchanged. Consider the example presented above again. When the prefix and the stem are combined into the word, neither of them goes through any phonological and orthographic changes.

On the other hand, Seneca also has fusional properties; this means that during the formation of some words, the combining morphemes can undergo phonological (and orthographical) changes. As an illustration, consider the following word in Seneca. When combining the four morphemes together, the masculine singular subject *hra*, the verb stem *k* and the *s* that marks habitual state do not undergo any changes; whereas the initial *i* is replaced with í to make sure that the verbs or verb phrases have at least two syllables (Chafe, 2015).

(2) íhrakis

i hra k s
it he eat HAB

He eats it.

In addition to Seneca, we include four Mexican indigenous languages from the Yuto-Aztecan language family (Baker, 1997) for our crosslinguisitic training experiments: Mexicanero (888 words), Nahuatl (1,123 words), Wixarika (1,385 words), and Yorem Nokki (1,063 words). The data for these languages contains morphological segmentation that was initially digitized from the book collections of *Archive of Indigenous Language* (Mexicanero (Una, 2001), Nahuatl (de Suárez, 1980), Wixarika (Gómez and López, 1999), Yorem Nokki (Freeze, 1989)). The data collection was carried out by the authors of Kann et al. (2018) based on the descriptions in their work, and their preprocessed data sets are publicly available. The four Yuto-Aztecan languages are also polysynthetic.

3 Related Work

The task of morphological segmentation has been cast in distinct ways in previous work. One line of

[1] https://senecalanguage.com/wp-content/uploads/Verb-Book-Vol.1.pdf

Language	Location	N. of speakers	Domain	Train	Dev	Test	Total
Seneca	Western New York	50	Grammar book	2,278	1,139	2,277	5,694
	Ontario		Informal sources	2,168	1,084	2,167	5,419
			Bible	-	-	-	8,588

Table 1: Descriptive information of the Seneca language and data.

research focuses on *surface segmentation* (Ruokolainen et al., 2016), while the other attends to *canonical segmentation* (Cotterell et al., 2016b). Both involve correctly decomposing a given word into distinct morphemes, which also typically includes words that stand alone as free morphemes.

Nevertheless, the two tasks differ in one key aspect: whether the combination [2] of the segmented morpheme sequence stays true to the initial orthography of the word. For surface segmentation, the answer is yes (e.g., Indonnesian *dihapus* → *di+hapus*). On the other hand, canonical segmentation sometimes involves the addition and/or deletion of characters from the surface form of the initial word, in order to capture phonological or orthographic characteristics of the component morphemes when uncombined. For example, the word *measurable* in English would be segmented as *measure* + *able*, recovering the orthographic *e* that was lost during word formation.

For surface segmentation, both supervised and unsupervised approaches have gained in popularity over the years. Within the realm of supervised methods, a large number of experiments have developed rule-based finite-state transducers (FST) (Kaplan and Kay, 1994) with weights usually determined by rich linguistic feature sets. The high functionality of hand-crafted FST for morphological analyses has been demonstrated for languages such as Persian (Amtrup, 2003), Finnish (Lindén et al., 2009), Semitic languages such as Tigrinya (Gasser, 2009) and Arabic (Beesley, 1996; Shaalan and Attia, 2012), as well as various African languages (Gasser, 2011). Other work has shifted to more data-driven machine learning techniques, including but not limited to memory-based learning (van den Bosch and Daelemans, 1999; Marsi et al., 2005), conditional random field models (CRF) (Cotterell et al., 2015; Ruokolainen et al., 2013, 2014), and convolutional networks (Sorokin and Kravtsova, 2018; Sorokin, 2019).

Unsupervised methods have perhaps enjoyed a longer history (Harris, 1955), with earlier studies relying on information-theoretic measures as indexes of character-level predictability, which were then used to determine morpheme boundaries (Hafer and Weiss, 1974). Later work such as Linguistica (Goldsmith, 2001) and Morfessor (Creutz and Lagus, 2002) applied the analyses of Minimum Description Length for morpheme induction (Rissanen, 1998; Poon et al., 2009). Goldwater et al. (2009) developed Bayesian generative models that would also take into account the context of individual words, which were able to simulate the process of how children learn to segment words given child-directed speech.

In contrast to surface segmentation, the problem of *canonical* segmentation has mainly been addressed with supervised methods. Cotterell et al. (2016b) extended a previous semi-CRF (Cotterell et al., 2015) for surface segmentation to jointly predict morpheme boundaries and orthographic changes, leading to improved results for German and Indonesian. With the same datasets, Kann et al. (2016) adopted character-based neural sequence models coupled with a neural reranker, presenting further improvement from Cotterell et al. (2016b). There has, however, been some unsupervised induction of canonical segmentation (see Hammarström and Borin (2011) for a thorough review). For instance, Dasgupta and Ng (2007) showed that certain spelling rules (e.g. insertion, deletion) derived heuristically from corpus frequency were able to handle orthographic changes during word formation. In comparison, Naradowsky and Goldwater (2009) provided a Bayesian model that formulate spelling rules probabilistically with character-level contextual information; the simultaneous learning process of both the rules and morpheme boundaries in turn boosted segmentation performance.

Although Seneca has fusional morphological features, meaning that certain morpheme boundaries within words are not necessarily clear-cut, the Seneca morphological data currently does not provide labeled canonical segmentation. We therefore focus on the task of surface segmentation.

[2] Here we used the term *combination* instead of *concatenation*, because surface segmentation is applicable to words with concatenative morphology as well as those with non-concatenative morphology.

4 Experiments

4.1 Data preprocessing

As mentioned in Section 2, the labeled words for Seneca came from both the verbal paradigm book by Bardeau (2007) and informal sources. We treated the two sources as separate domains and constructed a dataset for each. The number of morphemes per word on average in the grammar book is 3.87 (95% confidence intervals: (3.86, 3.88); see Section 4.4), which is slightly lower than that in the informal sources (4.12 (4.10, 4.13)). On the other hand, the number of unique morphemes is much higher in the data from the informal sources (N = 1,641) than that in the grammar book (N = 631). This difference in the amount of morphological variation between the two domains raises the expectation that with the same model architecture, morphological segmentation of the words from the informal sources is possibly more challenging.

For each data set, to construct the low-resource settings, we set the train/dev/test ratio to be 2:1:2, then randomly generated five splits for every dataset with this ratio (Gorman and Bedrick, 2019). [3] We used the first random split of both domains for model evaluation as well as selection of training settings; the setting(s) eventually selected would then be applied to data from each of the five random splits to test the stability of the model performance.

4.2 Evaluation design

We took advantage of the fact that the two data sets for Seneca came from different domains by investigating two experimental designs: evaluating with a development set versus evaluating with a *development domain*. The former carried out the standard practices. When building models for morphological segmentation of a particular domain, only the in-domain training set would be (part of) the training data for the models, along with possible addition of training data from the other domain or indigenous languages. The development set from the same domain would be used to evaluate models and the one(s) with the best performance would be selected (e.g. segmentation for the grammar book data using the development set of the grammar book for evaluation).

However, realistically development sets are luxuries to critically endangered languages (Kann et al., 2019). To help with the documentation of these languages more effectively, one would want to use as much training data as possible, ideally from the same domain or language. Yet acquiring more data for languages like Seneca, whether with or without manual annotations, faces extreme difficulty. It requires not only extensive time and financial resources, but also expertise from the very few native speakers left, most of whom are elders.

To increase the utility of the already-limited data for Seneca, we experimented with a second design of using a development domain for model evaluation. That is, for morphological segmentation of a particular domain, the new in-domain training data would be the concatenation of the initial training set along with the development set from the same domain. This new combination would be (part of) the training data for the models. In this case the development set of the other domain would then be applied instead to evaluate model performance (e.g. segmentation for the grammar book using the development set of the informal sources for evaluation). Again, the model(s) with the best performance on the development domain would be selected.

Comparing the two designs, taking into account the different configurations of the training data, it is possible that evaluation with a development domain would lead to different model architectures/settings being selected. On the other hand, it is also possible that the same model architecture or setting would be favored regardless of the particular design. In addition, because using a development domain essentially means that there is more in-domain training data, it remains to be seen whether this evaluation design would achieve better results when testing the stability of the model setting.

4.3 Model training

We experimented with three general settings: in-domain training, cross-domain training, and cross-linguistic training. For all settings, we adopted character-based sequence-to-sequence (seq2seq) recurrent neural network (RNN) (Elman, 1990) trained with OpenNMT (Klein et al., 2017). This model architecture has been previously demonstrated to perform well for polysynthetic indigenous languages (Kann et al., 2018).

In cases where applicable, we also compared the performance of the neural seq2seq models to unsupervised Morfessor [4] (Creutz and Lagus, 2002). In

[3] Data, code, and models are available at https://github.com/zoeyliu18/Seneca.

[4] In preliminary experiments, semi-supervised Morfes-

what follows, we describe the details of the seq2seq models in each training setting.

4.3.1 In-domain training

Naive baseline Our first baseline applied the default parameters in OpenNMT — an encoder-decoder long-short term memory model (LSTM) (Hochreiter and Schmidhuber, 1997) with the attention mechanism from Luong et al. (2015). All embeddings have 500 dimensions. Both the encoder and the decoder contain two hidden layers with 500 hidden units in each layer. Training was performed with SGD (Robbins and Monro, 1951) and a batch size of 64.

Abiding by our experimental designs, for all the baseline models, when evaluating with the development set, the in-domain training data came from just the training set. By contrast, when evaluating with the development domain, the in-domain training data was the concatenation of the training and the development sets.

Less naive baseline Going beyond the default settings in the first baseline, our second baseline experimented with different combinations of parameter settings and attention mechanisms (Bahdanau et al., 2015):

- RNN type: LSTM / GRU
- embedding dimesions: {128, 300, 500}
- hidden layers: {1, 2}
- hidden units: {128, 300, 500}
- batch size: {16, 32, 64}
- optimizer: SGD / ADADELTA (Zeiler, 2012)

These models were trained and evaluated in the same way as the first baseline. Based on results from either the development set or the development domain (after statistical tests; see Section 4.4), the model architecture that was selected was an attention-based encoder-decoder (Bahdanau et al., 2015), where the encoder is composed of a bidirectional GRU while the decoder consists of a unidirectional GRU. Both the encoder and the decoder have two hidden layers with 100 hidden states in each layer. All embeddings have 300 dimensions. Training was performed with ADADELTA and a batch size of 16.

sor (Kohonen et al., 2010) was also explored; yet the performance was worse than the unsupervised method. Thus we eventually chose the unsupervised variant for systematic comparisons with the seq2seq models.

4.3.2 Cross-domain training

With the model architecture of our *less naive* baseline, we turned to our cross-domain training experiments using four different methods.

Self-training The first method utilized self-training (McClosky et al., 2008) and resorted to the unlabeled words from the Bible, which were first automatically segmented with the second baseline model from in-domain training. These words were then added to the in-domain training data given each of the two evaluation designs (Section 4.2).

Multi-task learning The second method applied multi-task learning (Kann et al., 2018). In this case, in addition to the task of morphological segmentation, we added a new task where the training objective is to generate output that is identical to the input. In the seq2seq model, the decoder does not always generate every character in the input sequence, which prevents accurate morphological segmentation of the full word. Thus the ulterior goal of this additional task is simple yet important: helping the model learn to *copy*.

In particular, words from the in-domain training data were used for the segmentation task, while words from the Bible were used for mapping input to output. Every word in the eventual training data was appended with a task-specific input symbol. For instance, let X represent the task of morphological segmentation, Y the task of mapping input to output, the goal of the model is to jointly perform the following :

- ëwënötgëh + X → ë + wën + ötgëh
- oiwa' + Y → oiwa'

Transfer learning The third method adopts domain transfer learning. Consider morphological segmentation of the grammar book as an example. When using a development set, the in-domain training data, which includes only the training set of the grammar book, would be combined with *all* data from the informal sources. On the other hand, when using a development domain, the in-domain training data, which includes the training and development sets of the grammar book, would be concatenated with just the training and test sets from the informal sources.

Fine-tuning With the model trained from transfer learning, we fine-tuned it further with in-domain training data.

One point to note is when evaluating with a development domain, we expected that the model

trained with domain transfer learning (with fine-tuning) would yield the best results. However, these results would not be directly informative about whether this setting is indeed better than the others, the latter of which only included in-domain training data. Hence for this particular evaluation design, while we still carried out the domain transfer experiments for consistency, we selected models only based on the other training settings.

4.3.3 Cross-linguistic training

In order to examine whether data from other polysynthetic languages would improve model performance, we carried out cross-linguistic training with three different settings: multi-task learning, transfer learning (Kann et al., 2018), and fine-tuning. These settings are similar to those in cross-domain training, except that the data from the four Mexican languages was used as additional training data instead of the Bible or out-of-domain data.

4.4 Metrics

Three measures were computed as indexes of model performance (Cotterell et al., 2016a; van den Bosch and Daelemans, 1999): full form accuracy, morpheme F1, and average Levenshtein distance (Levenshtein, 1966). Significance testing of each metric was conducted with bootstrapping (Efron and Tibshirani, 1994). As an illustration, take full form accuracy as an example. After applying a model to the development set (or domain) with a total of N words, we: (1) randomly selected N words from the development set with replacement; (2) calculated the full form accuracy of the selected sample; (3) repeated step (1) and (2) for 10,000 iterations, which yielded an empirical distribution of full form accuracy; (4) measured the mean and the 95% confidence interval (CI) of the empirical distribution.

5 Results

5.1 Evaluation with development set

For evaluation, we considered a training setting to be better than another based on at least one of the three metrics calculated. As presented in Table 2, when evaluating with the development set, it appears that for the grammar book, the simple less naive baseline with careful parameter tuning is able to yield excellent performance, while other more complicated training configurations such as including additional out-of-domain data do not lead to further improvement (no significant differences in the results). Therefore we chose the less naive baseline from in-domain training for the final testing given its simplicity and average score for each of the three metrics.

By contrast, with the same training settings, the models show weaker performance for informal sources. This corresponds to our initial expectation that due to the higher number of unique morphemes in informal sources, accurately labeling the boundaries of these morphemes would be comparatively more challenging. Similar to results for the grammar book, none of the other training configurations seems to significantly surpass the two baselines. With that being said, we selected the cross-linguistic training with multi-task learning for the final testing, again because it has the best average score for each of the three measures.

5.2 Evaluation with development domain

On the other hand, when evaluating with the development domain, as shown in Table 3, almost all other training configurations appear to be better than the two baselines, a pattern that holds for data from the grammar book as well as that from the informal sources. When compared to the two baselines, while the other settings do not show significant improvement in terms of accuracy or F1 score, the average Levenshtein distance is shorter when the models are trained with multi-task learning and/or additional cross-linguistic data. Given the results, for both the grammar book and the informal sources, we selected cross-domain multi-task learning as the setting for final model testing.

Combining the results from Table 2 and Table 3 together, it appears that regardless of the particular evaluation design, in any of the settings where unsupervised Morfessor is applicable (Creutz and Lagus, 2002), the neural encoder-decoder models consistently yielded significantly better performance in relation to all three measures. This observation also speaks to previous findings from Kann et al. (2018), except that they adopted semi-supervised variants of Morfessor.

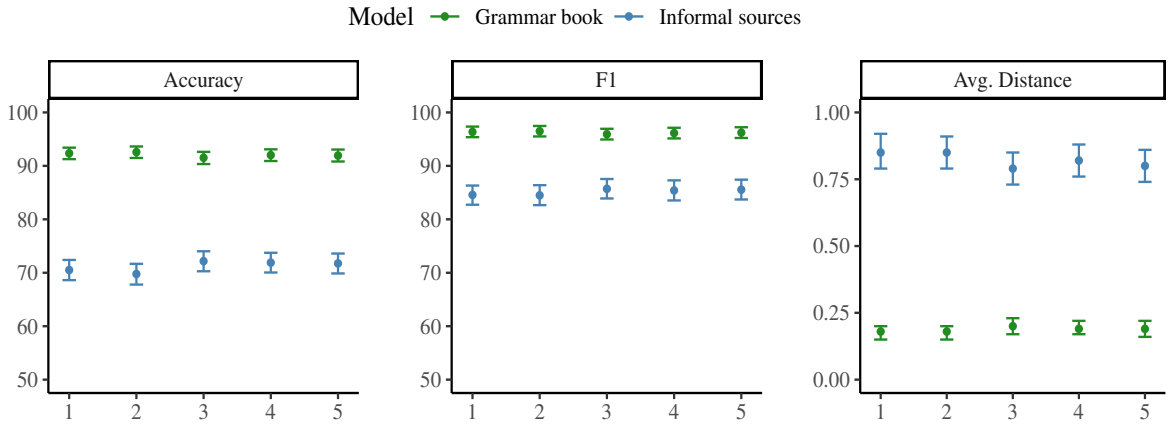
Comparing the segmentation results from the seq2seq models to those from Morfessor, overall there does not seem to be aspects where the latter systematically falls short, in the sense that the segmentation patterns by Morfessor are more or less "all over the place". One potential explanation lies in the fact that in both our data sets, the majority of

Grammar book	Models	Accuracy	F1	Avg. Distance	better than Morfessor?	Selected?
In-domain	*naive baseline*	86.03	93.10	0.39	Yes	
	less naive baseline	91.92	95.96	0.21	Yes	✓
Cross-domain	*self-training*	89.98	95.04	0.26	Yes	
	multi-task learning	91.38	95.78	0.21	Yes	
	transfer learning	86.02	92.54	0.39	Yes	
	fine-tuning	88.68	94.21	0.29		
Cross-linguistic	*multi-task learning*	91.06	95.50	0.22	Yes	
	transfer learning	90.00	95.15	0.24	Yes	
	fine-tuning	90.16	95.22	0.24		
Informal sources						
In-domain	*naive baseline*	69.99	84.47	0.96	Yes	
	less naive baseline	71.38	85.27	0.86	Yes	
Cross-domain	*self-training*	70.05	84.74	0.87	Yes	
	multi-task learning	72.04	85.38	0.83	Yes	
	transfer learning	67.42	82.50	0.98	Yes	
	fine-tuning	69.27	83.79	0.92		
Cross-linguistic	*multi-task learning*	73.51	86.04	0.78	Yes	✓
	transfer learning	70.95	85.19	0.83	Yes	
	fine-tuning	71.39	85.35	0.82		

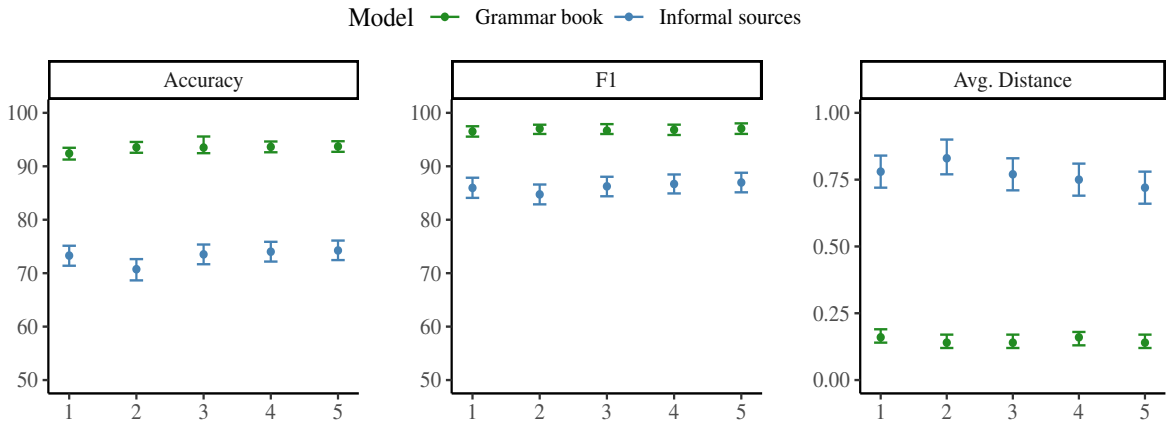
Table 2: Model training and evaluation with **the development set**. The value of each metric for every model was compared to those of the two baselines; boldface indicates significant differences from **both baselines**, derived by comparing their respective 95% CI after bootstrapping. Selected training setting for model testing is checkmarked.

Grammar book	Models	Accuracy	F1	Avg. Distance	better than Morfessor?	Selected?
In-domain	*naive baseline*	11.43	40.32	5.90	Yes	
	less naive baseline	12.35	40.77	4.01	Yes	
Cross-domain	*self-training*	13.38	42.96	3.77	Yes	
	multi-task learning	14.66	42.97	**3.24**	Yes	✓
Cross-linguistic	*multi-task learning*	12.54	41.63	**3.28**	Yes	
	transfer learning	15.12	40.89	**3.40**	Yes	
	fine-tuning	15.52	41.15	**3.40**		
Informal sources						
In-domain	*naive baseline*	10.18	44.16	4.58	Yes	
	less naive baseline	12.97	45.38	3.66		
Cross-domain	*self-training*	12.92	45.08	**3.31**	Yes	
	multi-task learning	16.59	47.79	**2.97**	Yes	✓
Cross-linguistic	*multi-task learning*	14.65	45.91	**3.15**	Yes	
	transfer learning	13.61	45.07	**3.07**	Yes	
	fine-tuning	13.61	45.24	**3.06**		

Table 3: Model training and evaluation with **the development domain**. The value of each metric for every model was compared to those of the two baselines; boldface indicates significant differences from **both baselines**, derived by comparing their respective 95% CI after bootstrapping. Selected training setting for model testing is checkmarked.



(a) using the development set for evaluation



(b) using the development domain for evaluation

Figure 1: Model testing results given different evaluation designs; error bars indicate 95% CI after bootstrapping.

the words have a frequency of one (95.28% for the grammar book; 95.57% for the informal sources). On the other hand, successful segmentation by unsupervised Morfessor relies heavily on the frequency of a given word and accordingly the number of overlapping or common morphemes shared by different words, whether the occurrence frequency information was computed from the training data or from additional unlabeled data. In addition to the complex morphological features of Seneca and the high frequency of unique morphemes in the two data sets used in our experiments, the Bible dataset, despite containing more unlabeled words, is still relatively small (N = 8,588), and thus is not especially useful for deriving frequency estimates.

5.3 Testing

For both the grammar book and the informal sources, we tested the stability of the selected model settings across the five random splits (Section 4.1). With each random split, we trained a model following the selected setting for each of the evaluation designs; the model was then applied to the test set of the random split.

Based on Figure 1, within each evaluation design, the test performance of the model setting is stable across the random splits. Morphological segmentation of data from the grammar book was able to achieve consistently better results than that for the informal sources. Regardless of the data source, while there does not appear to be significant differences in model performance between the two evaluation designs, comparing to using a development set, evaluating with a development domain led to slight improvement of average scores for each of the three metrics.

6 Conclusions and Future Work

We have investigated morphological segmentation for Seneca, an indigenous Native American language with highly complex morphological characteristics. In a series of in-domain, cross-domain, and cross-linguistic training settings, the results demonstrate that neural seq2seq models are quite effective at correctly labeling morpheme boundaries, at least at the surface level. With the two evaluation designs explored here, the model settings were able to achieve above 96% F1 score for data from the grammar book, and above 85% for the informal sources.

Many of the languages indigenous to North America are as endangered as Seneca and have available resources comparable in both size and scope to those used in the current work. Our thorough investigation of how to effectively integrate these limited and varied resources can potentially serve as a model for other community-driven collaborations to document endangered languages for future generations, and to produce materials suitable for language immersion and revitalization. For our future work, in addition to refining and improving our models, we also plan to explore the utility of morphological segmentation for improving language modeling in ASR. This would be able to support transcription of both archival recordings and new recordings captured by community members involved in language revitalization projects.

Acknowledgements

We are grateful for the cooperation and support of the Seneca Nation of Indians. This material is based upon work supported by the National Science Foundation under Grant No. 1761562. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Mohamed Afify, Ruhi Sarikaya, Hong-Kwang Jeff Kuo, Laurent Besacier, and Yuqing Gao. 2006. On the use of morphological analysis for dialectal arabic speech recognition. In *Ninth International Conference on Spoken Language Processing*.

Alexandra Y Aikhenvald et al. 2007. Typological distinctions in word-formation. *Language typology and syntactic description*, 3:1–65.

Jan W Amtrup. 2003. Morphology in machine translation systems: Efficient integration of finite state transducers and feature structure descriptions. *Machine Translation*, 18(3):217–238.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Mark C Baker. 1997. Complex predicates and agreement in polysynthetic languages. *Complex predicates*, pages 247–288.

Phyllis E. Wms. Bardeau. 2007. *The Seneca Verb: Labeling the Ancient Voice*. Seneca Nation Education Department, Cattaraugus Territory.

Kenneth R Beesley. 1996. Arabic finite-state morphological analysis and generation. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Wallace L Chafe. 2015. *A Grammar of the Seneca Language*, volume 149. University of California Press.

Ann Clifton and Anoop Sarkar. 2011. Combining morpheme-based machine translation with post-processing morpheme prediction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 32–42, Portland, Oregon, USA. Association for Computational Linguistics.

Ryan Cotterell, Arun Kumar, and Hinrich Schütze. 2016a. Morphological segmentation inside-out. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2325–2330, Austin, Texas. Association for Computational Linguistics.

Ryan Cotterell, Thomas Müller, Alexander Fraser, and Hinrich Schütze. 2015. Labeled morphological segmentation with semi-Markov models. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 164–174, Beijing, China. Association for Computational Linguistics.

Ryan Cotterell, Tim Vieira, and Hinrich Schütze. 2016b. A joint model of orthography and morphological segmentation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 664–669, San Diego, California. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 21–30. Association for Computational Linguistics.

Sajib Dasgupta and Vincent Ng. 2007. High-performance, language-independent morphological segmentation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 155–163, Rochester, New York. Association for Computational Linguistics.

Yolanda Lastra de Suárez. 1980. *Náhuatl de Acaxochitlán (Hidalgo)*. El Colegio de México.

Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. CRC press.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Ray A Freeze. 1989. *Mayo de Los Capomos, Sinaloa*. El Colegio de México.

Michael Gasser. 2009. Semitic morphological analysis and generation using finite state transducers with feature structures. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 309–317.

Michael Gasser. 2011. Hornmorpho: a system for morphological processing of amharic, oromo, and tigrinya. In *Conference on Human Language Technology for Development, Alexandria, Egypt*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. In *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational linguistics*, 27(2):153–198.

Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. 2009. A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Paula Gómez and Paula Gómez López. 1999. *Huichol de San Andrés Cohamiata, Jalisco*. El Colegio de México.

Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791, Florence, Italy. Association for Computational Linguistics.

Joseph H Greenberg. 1960. A quantitative approach to the morphological typology of language. *International journal of American linguistics*, 26(3):178–194.

Margaret A Hafer and Stephen F Weiss. 1974. Word segmentation by letter successor varieties. *Information storage and retrieval*, 10(11-12):371–385.

Harald Hammarström and Lars Borin. 2011. Unsupervised learning of morphology. *Computational Linguistics*, 37(2):309–350.

Zellig S Harris. 1955. From phoneme to morpheme. *Language*, 31(2):190–222.

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

Robbie Jimerson and Emily Prud'hommeaux. 2018. ASR for documenting acutely under-resourced indigenous languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Katharina Kann, Kyunghyun Cho, and Samuel R. Bowman. 2019. Towards realistic practices in low-resource natural language processing: The development set. 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 3342–3349, Hong Kong, China. Association for Computational Linguistics.

Katharina Kann, Ryan Cotterell, and Hinrich Schütze. 2016. Neural morphological analysis: Encoding-decoding canonical segments. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 961–967, Austin, Texas. Association for Computational Linguistics.

Katharina Kann, Jesus Manuel Mager Hois, Ivan Vladimir Meza-Ruiz, and Hinrich Schütze. 2018. Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. 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 47–57, New Orleans, Louisiana. Association for Computational Linguistics.

Ronald M Kaplan and Martin Kay. 1994. Regular models of phonological rule systems. *Computational linguistics*, 20(3):331–378.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Oskar Kohonen, Sami Virpioja, and Krista Lagus. 2010. Semi-supervised learning of concatenative morphology. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 78–86.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Krister Lindén, Tommi Pirinen, et al. 2009. Weighted finite-state morphological analysis of Finnish compounding with hfst-lexc. In *NEALT Proceedings Series*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Erwin Marsi, Antal van den Bosch, and Abdelhadi Soudi. 2005. Memory-based morphological analysis generation and part-of-speech tagging of Arabic. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 1–8, Ann Arbor, Michigan. Association for Computational Linguistics.

David McClosky, Eugene Charniak, and Mark Johnson. 2008. When is self-training effective for parsing? In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 561–568, Manchester, UK. Coling 2008 Organizing Committee.

Barbra A Meek. 2012. *We are our language: An ethnography of language revitalization in a Northern Athabaskan community*. University of Arizona Press.

Jason Naradowsky and Sharon Goldwater. 2009. Improving morphology induction by learning spelling rules. In *Proceedings of the 21st international jont conference on Artifical intelligence*, pages 1531–1536.

Karthik Narasimhan, Damianos Karakos, Richard Schwartz, Stavros Tsakalidis, and Regina Barzilay. 2014. Morphological segmentation for keyword spotting. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 880–885, Doha, Qatar. Association for Computational Linguistics.

Graham Neubig, Shruti Rijhwani, Alexis Palmer, Jordan MacKenzie, Hilaria Cruz, Xinjian Li, Matthew Lee, Aditi Chaudhary, Luke Gessler, Steven Abney, Shirley Anugrah Hayati, Antonios Anastasopoulos, Olga Zamaraeva, Emily Prud'hommeaux, Jennette Child, Sara Child, Rebecca Knowles, Sarah Moeller, Jeffrey Micher, Yiyuan Li, Sydney Zink, Mengzhou Xia, Roshan S Sharma, and Patrick Littell. 2020. A summary of the first workshop on language technology for language documentation and revitalization. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 342–351, Marseille, France. European Language Resources association.

Hoifung Poon, Colin Cherry, and Kristina Toutanova. 2009. Unsupervised morphological segmentation with log-linear models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 209–217.

Jorma Rissanen. 1998. *Stochastic complexity in statistical inquiry*, volume 15. World scientific.

Herbert Robbins and Sutton Monro. 1951. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Teemu Ruokolainen, Oskar Kohonen, Kairit Sirts, Stig-Arne Grönroos, Mikko Kurimo, and Sami Virpioja. 2016. A comparative study of minimally supervised morphological segmentation. *Computational Linguistics*, 42(1):91–120.

Teemu Ruokolainen, Oskar Kohonen, Sami Virpioja, and Mikko Kurimo. 2013. Supervised morphological segmentation in a low-resource learning setting using conditional random fields. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 29–37, Sofia, Bulgaria. Association for Computational Linguistics.

Teemu Ruokolainen, Oskar Kohonen, Sami Virpioja, and Mikko Kurimo. 2014. Painless semi-supervised morphological segmentation using conditional random fields. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 84–89, Gothenburg, Sweden. Association for Computational Linguistics.

Wolfgang Seeker and Özlem Çetinoğlu. 2015. A graph-based lattice dependency parser for joint morphological segmentation and syntactic analysis. *Transactions of the Association for Computational Linguistics*, 3:359–373.

Khaled Shaalan and Mohammed Attia. 2012. Handling unknown words in Arabic FST morphology. In *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*, pages 20–24, Donostia–San Sebastián. Association for Computational Linguistics.

Alexey Sorokin. 2019. Convolutional neural networks for low-resource morpheme segmentation: baseline or state-of-the-art? In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 154–159.

Alexey Sorokin and Anastasia Kravtsova. 2018. Deep convolutional networks for supervised morpheme segmentation of Russian language. In *Conference on Artificial Intelligence and Natural Language*, pages 3–10. Springer.

John Sylak-Glassman. 2016. The composition and use of the universal morphological feature schema (unimorph schema). *Johns Hopkins University*.

Bao Thai, Robert Jimerson, Dominic Arcoraci, Emily Prud'hommeaux, and Raymond Ptucha. 2019. Synthetic data augmentation for improving low-resource asr. In *2019 IEEE Western New York Image and Signal Processing Workshop (WNYISPW)*, pages 1–9. IEEE.

Bao Thai, Robert Jimerson, Raymond Ptucha, and Emily Prud'hommeaux. 2020. Fully convolutional ASR for less-resourced endangered languages. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 126–130, Marseille, France. European Language Resources association.

Canger Una. 2001. *Mexicanero de la sierra madre occidental.* El Colegio de México.

Antal van den Bosch and Walter Daelemans. 1999. Memory-based morphological analysis. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 285–292, College Park, Maryland, USA. Association for Computational Linguistics.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.