Proceedings of SIGUL2022 @LREC2022, pages 155–160
Marseille, 24-25 June 2022

Evaluating Unsupervised Approaches to Morphological Segmentation for Wolastoqey

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

Abstract

Finite-state approaches to morphological analysis have been shown to improve the performance of natural language processing systems for polysynthetic languages, in-which words are generally composed of many morphemes, for tasks such as language modelling (Schwartz et al., 2020). However, finite-state morphological analyzers are expensive to construct and require expert knowledge of a language's structure. Currently, there is no broad-coverage finite-state model of morphology for Wolastoqey, also known as Passamaquoddy-Maliseet, an endangered low-resource Algonquian language. As this is the case, in this paper, we investigate using two unsupervised models, MorphAGram and Morfessor, to obtain morphological segmentations for Wolastoqey. We train MorphAGram and Morfessor models on a small corpus of Wolastoqey words and evaluate using two annotated datasets. Our results indicate that MorphAGram outperforms Morfessor for morphological segmentation of Wolastoqey.

Keywords: Morphological segmentation, unsupervised morphology, low-resource languages

1. Introduction

Wolastoqey is an Indigenous language spoken in parts of what are now the provinces of New Brunswick and Quebec, Canada, and the state of Maine, United States. This language is often referred to as Passamaquoddy-Maliseet, with Passamaquoddy and Maliseet being two dialects of this language. Many speakers of the Maliseet dialect in the communities where the authors of this paper live and work refer to their language as Wolastoqey. We therefore use the term *Wolastoqey* (as opposed to Passamaquoddy-Maliseet) in this paper.

Wolastoqey is a polysynthetic eastern Algonquian language. It is endangered, with only roughly 300 remaining first language speakers in Canada (Statistics Canada, 2017). It is a low-resource language with no large corpora or annotated datasets available. There is, however, the Passamaquoddy-Maliseet Dictionary (Francis and Leavitt, 2008). This Wolastoqey–English dictionary provides English definitions for roughly $19k$ Wolastoqey headwords. A version of this dictionary is available online.[1]

Relatively little prior computational work has considered Wolastoqey. Farber (2015) presents a preliminary finite-state model of Passamaquoddy-Maliseet noun morphology. Bear and Cook (2021) propose a cross-lingual Wolastoqey–English definition modelling system which generates English definitions for Wolastoqey words. They show that, for this definition modelling task, sub-word representations from byte-pair encoding (Sennrich et al., 2016) can be used to overcome the limitations of not having a large monolingual Wolastoqey corpus available for learning Wolastoqey word representations. Bear and Cook (2022) show that English definitions for Wolastoqey words can be used to form Wolastoqey word representations that encode syntactic and semantic information.

Morphological analysis is particularly important for building language technology and natural language processing systems for morphologically-rich languages. For example, Bowers et al. (2017) develop a morphological parser for the Odawa dialect of Ojibwe (also an Algonquian language) and discuss applications of this parser for building language technology such as morphologically-aware dictionary search to help a dictionary user to find a lemma from an inflected form and spelling correction. A Wolastoqey morphological analyzer could similarly enable such language technologies for this language. Finite-state morphology has also been shown to give improvements in language modelling for polysynthetic languages (Schwartz et al., 2020). Language models are a key component for many NLP systems. As such, a Wolastoqey morphological analyzer could support the development of future applications such as text prediction.

Finite state morphological analyzers have been developed for several Algonquian languages including Plains Cree (Snoek et al., 2014), Odawa (Bowers et al., 2017), and Arapaho (Kazeminejad et al., 2017). However, other than the preliminary work of Farber (2015) on noun morphology, there is currently no broad coverage finite state morphological analyzer for Wolastoqey.

In the absence of a finite state morphological analyzer for Wolastoqey, in this paper, we consider unsupervised approaches to morphological segmentation. MorphAGram (Eskander et al., 2020) is an unsupervised approach to morphological segmentation based on adaptor grammars, models that generalize probabilistic context-free grammars by introducing depen-

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

dencies between successive uses of rewrite rules. It has recently been shown to outperform other unsupervised approaches to morphological segmentation on a range of languages, including polysynthetic languages. In this paper, we evaluate MorphAGram on Wolastoqey, and compare it to Morfessor (Smit et al., 2014), an unsupervised morphological segmentation model that defines a segmentation vocabulary using minimum description length as a training objective. We find that MorphAGram also outperforms Morfessor for Wolastoqey.

The rest of the paper is organized as follows. In Section 2 we describe our experimental setup including the models considered, the training and evaluation datasets, and the evaluation metric. We present results for MorphAGram and Morfessor in Section 3. In Section 4 we summarize our findings and identify directions for future work.

2. Experimental Setup

In this section we describe the settings of MorphAGram and Morfessor used in our experiments, the training and evaluation data, and the evaluation metrics we use.

2.1. MorphAGram

To run our experiments with MorphAGram, we use the implementation of MorphAGram published by Eskander et al. (2020). This implementation requires an off the shelf adaptor-grammar sampler to train; we use the recommended adapter-grammar sampler.[2] To train our MorphAGram models, we use the same training parameters as the original paper as described in the source code of the implementation.[3]

As we wish to evaluate the performance of MorphAGram on Wolastoqey, we first must identify the best performing grammar for this language. For this, we consider running preliminary experiments in which we train multiple MorphAGram models using the grammars considered by Eskander et al. (2020). We evaluate the performance of each grammar on a small dataset of morphologically segmented words from the Passamaquoddy-Maliseet Dictionary (the PMLP dataset described in 2 4). In these preliminary experiments, we observed that a grammar consisting of prefixes, stems and suffixes, referred to PrStSu in the original paper, performed the best. We therefore choose to focus on this grammar, as well as the best performing grammar from the original paper, which, in-addition to prefixes, stems and suffixes, includes submorphs. This grammar is referred to as PrStSu + SM.

We choose to run our experiments both in a language-independent and scholar-seeded configuration. To train our models in a scholar-seeded setup, we seed our grammars using preverbs from the Passamaquoddy-Maliseet Dictionary. In total, we seed our scholar-seeded grammars with 813 preverbs.[4]

2.2. Morfessor

To establish a baseline for comparison, we train a Morfessor 2.0 (Smit et al., 2014) model on the same datasets used to train our MorphAGram models. For this we use the implementation of Morfessor 2.0 available in the python Morfessor library.[5] The Morfessor model used in our experiments is trained using the default training parameters on the types that occur in our training dataset.

2.3. Training Data

To construct the training dataset used in our experiments we use contents from the Passamaquoddy-Maliseet Dictionary. In addition to English definitions for Wolastoqey headwords, this dictionary includes parallel Wolastoqey–English example sentences. As we require a list of words to train our morphological segmentation models, we define our training dataset as the set of unique types that occur in the Wolastoqey example sentences of each dictionary entry. We choose to use the types that occur in the dictionary example sentences instead of the set of dictionary headwords, as all verb headwords are given in a third-person present-tense form, meaning many morphemes associated with particular inflected forms would not occur in the training data.

To obtain a list of types from our Wolastoqey sentences, we first tokenize each sentence using a regular expression tokenizer from NLTK (Bird and Loper, 2004). We define a token as a contiguous string of alphanumeric characters, underscores, hyphens and apostrophes. As many example sentences code-switch with English and thus contain English words, we remove all English words from our dataset using an English word list available in NLTK. Using this approach, we obtain a set of $30.1k$ unique types to train our models from $18.5k$ example sentences, containing a total of $147k$ tokens.

As both Morfessor and MorphAGram are unsupervised approaches to morphological segmentation, we choose to evaluate our models in a transductive setup in which words the model will be evaluated on (but not their gold-standard segmentations) are included in the training data. Given new unknown words to segment, the models could be simply retrained to obtain segmentations for them. Operating under this assumption, for each of our experiments, we add all words from the evaluation set (described below) to the training data.

[2] https://web.science.mq.edu.au/~mjohnson/Software.htm

[3] https://github.com/rnd2110/MorphAGram

[4] Many preverbs are listed in the Passamoquoddy-Maliseet Dictionary as headwords and as such can easily be identified to use in a scholar-seeded setting. In future work we intend to also consider including common suffixes in the scholar-seeded setting.

[5] https://github.com/aalto-speech/morfessor

2.4. Evaluation Datasets

For evaluation, we compare the output of MorphAGram and Morfessor to gold standard segmentations. We use two segmentation datasets for evaluation, one obtained from information available on the Passamaquoddy-Maliseet Language Portal, and the other from a morphologically-annotated sample text (Leavitt, 1996, 5.4).

The Passamaquoddy-Maliseet Language Portal includes word-building examples to help teach learners how words are formed.[6] These examples include information about morphological segmentation. We use all of the available examples to form a dataset for evaluation. The resulting dataset, which we refer to as PMLP, contains segmentations for 30 Wolastoqey words, composed of an average of 4.23 morphemes per word.

We build a second evaluation dataset from a morphologically-annotated sample text (Leavitt, 1996, 5.4). In this text, the morphological segmentation of each word is shown. We manually transcribe this sample text to create an additional evaluation dataset. This dataset, which we refer to as LEAVITT-1996, is composed of segmentations for 102 unique words (types), consisting of an average of 2.32 morphemes per word. LEAVITT-1996 is derived from running text. As such, it includes words corresponding to all parts-of-speech, including mono-morphemic particles and preverbs. This is in contrast to PMLP in which all instances in the dataset consist of multiple morphemes. Particles and preverbs can, however, be easily identified using a wordlist. As such, we are particularly interested in how a morphological segmenter performs on other parts-of-speech. We therefore construct a version of LEAVITT-1996 in which particles and preverbs are removed. We refer to this dataset as LEAVITT-1996-FILTERED. This results in a dataset consisting of segmentations for 71 words, being composed on average of 2.89 morphemes. For evaluations using LEAVITT-1996-FILTERED, we also remove particles and preverbs from the training data. This reduces the training data to 29.5k types as 624 particles and preverbs are removed from the training data.

2.5. Evaluation Metrics

A range of evaluation metrics have been considered for evaluating unsupervised morphological analyzers including boundary evaluations and morpheme assignment approaches such as EMMA-2 (Virpioja et al., 2011). In the case that both the predicted analysis and gold-standard are segmentations, Virpioja et al. (2011) find that boundary evaluations are appropriate. In our experimental setup both the predicted analyses and gold-standard annotations are segmentations, and so we use boundary precision-recall (BPR) for evaluation. BPR is a metric based on the precision, recall and F1 score of predicted segmentation splits.

[6] https://pmportal.org/word-building

3. Results

We report results for MorphAGram and Morfessor on each dataset in Table 1. For MorphAGram we consider a grammar with prefixes, stems, and suffixes (PrStSu) and the same grammar additionally with submorphs (PrStSu + SM). We consider each grammar in both a standard language-independent setting (Std.) and a scholar-seeded setting in which the model is seeded with knowledge of preverbs (Sch.). Results for MorphAGram approaches are averaged across ten runs with different random seeds.

We first consider results on PMLP (shown in the top panel of the Table 1). Focusing on F1, we observe that all MorphAGram approaches considered outperform the Morfessor baseline. This is inline with the findings of Eskander et al. (2020) that MorphAGram improves over Morfessor. Among the MorphAGram approaches considered we observe that the best approach is Std. PrStSU, i.e., a model without submorphs that does not use scholar seeding. We find that both approaches that do not use submorphs outperform those that do, and that using scholar seeding leads to a reduction in performance.

We now turn to consider results on LEAVITT-1996 (middle panel of Table 1). Focusing again on F1, we observe that for this dataset, not all MorphAGram models outperform the Morfessor baseline. In particular, only models that incorporate submorphs (indicated with +SM) outperform Morfessor. In contrast to experiments on PMLP, here we observe that both approaches that incorporate submorphs outperform those that do not.

We further see mixed results here for scholar seeding. In particular, scholar seeding gives a small improvement for models that do not use submorphs, but does not give improvements when submorphs are included. The best results on this dataset use submorphs and no scholar seeding (i.e., Std. PrStSu + SM). The inconsistent behaviour of scholar seeding could possibly be attributed to the fact that we only use prefixes as seeds in our experiments, and do not use stems or suffixes as seeds. Additionally providing stems and suffixes as part of the scholar seeding could potentially lead to improvements. However, the finding that scholar-seeding does not lead to uniform benefits is not inconsistent with Eskander et al. (2020) who find that scholar-seeding did not improve performance on some languages.

In the PMLP evaluation, all instances consist of multiple morphemes. In contrast, for LEAVITT-1996, the instances are drawn from running text and include many particles and preverbs (the latter of which are in certain cases written as separate words) which are mono-morphemic. In preliminary investigations we observed that MorphAGram over-segmented many of these monomorphemic forms, which seems to have contributed to the relatively low precision of MorphAGram approaches compared to Morfessor on LEAVITT-

PMLP

Grammar	P		R		F1	
Morfessor	0.678		0.377		0.485	
Std. PrStSu	0.619	(0.026)	**0.623**	(0.021)	**0.621**	(0.021)
Std. PrStSu + SM	0.736	(0.021)	0.504	(0.027)	0.598	(0.024)
Sch. PrStSu	0.644	(0.022)	0.571	(0.030)	0.605	(0.025)
Sch. PrStSu + SM	**0.738**	(0.031)	0.466	(0.025)	0.571	(0.026)

LEAVITT-1996

Morfessor	**0.710**		0.588		0.643	
Std. PrStSu	0.417	(0.022)	**0.800**	(0.022)	0.548	(0.023)
Std. PrStSu + SM	0.611	(0.021)	0.757	(0.017)	**0.676**	(0.018)
Sch. PrStSu	0.450	(0.025)	0.737	(0.019)	0.559	(0.022)
Sch. PrStSu + SM	0.605	(0.025)	0.747	(0.016)	0.668	(0.017)

LEAVITT-1996-FILTERED

Morfessor	0.668		0.452		0.539	
Std. PrStSu	0.544	(0.025)	**0.668**	(0.021)	0.599	(0.022)
Std. PrStSm + SM	**0.772**	(0.032)	0.616	(0.022)	**0.685**	(0.022)
Sch. PrStSm	0.630	(0.022)	0.617	(0.019)	0.623	(0.018)
Sch. PrStSm + SM	0.763	(0.019)	0.599	(0.020)	0.671	(0.016)

Table 1: Boundary precision, recall and F1 scores on each dataset for MorphAGram and a Morfessor 2.0 baseline. The standard deviation for these evaluation metrics for MorphAGram is shown in parentheses. The best results for each method, on each dataset, are shown in boldface.

Word	Approach	Segmentation
alitahasuwinuwok	Gold standard	ali+tahas+uwin+uwok
	MorphAGram	ali+tahas+uwin+uwok
	Morfessor	al+itahasu+winuwok
kpeciptulonen	Gold standard	k+peci+pt+ul+on+ en
	MorphAGram	k+pecip+t+ul+on+en
	Morfessor	kpeci+ptul+onen
wicihtaqik	Gold standard	wici+ht+aq+ik
	MorphAGram	wi+ci+ht+a+qik
	Morfessor	wici+htaq+ik

Table 2: The segmentations for the gold standard, MorphAGram, and Morfessor for three words in LEAVITT-1996.

1996. For example, MorphAGram segments the mono-morphemic preverb *cuwi* as c+uwi while Morfessor does not segment this word. These findings led us to consider a further evaluation on LEAVITT-1996-FILTERED in which particles and preverbs are excluded from the evaluation.

Results for LEAVITT-1996-FILTERED are shown in the bottom panel of Table 1. In this evaluation, as for the evaluation on PMLP, all MorphAGram methods outperform the Morfessor baseline. For this evaluation the results follow a similar pattern to those on the full LEAVITT-1996 dataset. Including submorphs gives improvements, while the results for scholar seeding are mixed; the best results are again obtained using submorphs and no scholar seeding (i.e., Std. PrStSu + SM).

Further comparing the results between LEAVITT-1996 and LEAVITT-1996-FILTERED, we observe that Morfessor performs notably worse on the latter. This suggests that Morfessor performs well at (not) segmenting mono-morphemic words such as particles and preverbs. Such words can, however, be easily identified using a wordlist. We further observe that each MorphAGram approach achieves higher precision on LEAVITT-1996-FILTERED than on LEAVITT-1996. This finding is inline with the observation that MorphAGram oversegments monomorphemic items, which are included in LEAVITT-1996 but not LEAVITT-1996-FILTERED.

Table 2 shows some examples of the segmentations produced by MorphAGram and Morfessor. For *alitahasuwinuwok* ('the wise men') MorphAGram produces the same segmentation as the gold standard, while none of the boundaries predicted by Morfessor are correct. In the case of *kpeciptulonen* ('constant battles') Mor-

phAGram produces an almost correct segmentation, but one boundary is incorrectly identified. For Morfessor, all predicted boundaries are correct, but recall is poor in that some boundaries are not predicted. For *wicihtaqik* ('make jointly') MorphAGram makes several errors, while Morfessor only fails to identify one boundary.

4. Conclusions

A morphological analyzer can be leveraged to give improvements for NLP tasks such as language modelling for polysynthetic languages. There is, however, currently no broad-coverage morphological analyzer for Wolastoqey. In this paper we therefore considered unsupervised approaches to morphological segmentation for Wolastoqey. MorphAGram has previously been shown to outperform Morfessor on polysynthetic languages. In this paper we evaluated MorphAGram and Morfessor and showed that this is also the case for Wolastoqey.

In future work, we intend to develop a finite-state morphological analyzer for Wolastoqey. Such a system could subsequently be leveraged to train a neural morphological analyzer with broader coverage (Micher, 2017; Lane and Bird, 2020). We are further interested in extrinsic evaluation of the segmentations produced by MorphAGram and leveraging them in applications. For example, we intend to consider whether cross-lingual Wolastoqey–English definition modelling could be improved by replacing BPE-based subword representations with segmentations from MorphAGram in the approach of Bear and Cook (2021). We are further interested in applications of morphological segmentations for semi-automated lexicography. For example, dictionaries of other Algonquian languages include entries for stems, roots, and affixes (Frantz and Russell, 2017). We are interested in whether MorphAGram segmentations could be leveraged to help lexicographers to add similar entries to a Wolastoqey dictionary.

5. Bibliographical References

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.

Bear, D. and Cook, P. (2022). Leveraging a bilingual dictionary to learn wolastoqey word representations. To appear in *Proceedinsg of LREC 2022*.

Bird, S. and Loper, E. (2004). NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain, July. Association for Computational Linguistics.

Bowers, D., Arppe, A., Lachler, J., Moshagen, S., and Trosterud, T. (2017). A morphological parser for odawa. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–9, Honolulu, March. Association for Computational Linguistics.

Eskander, R., Callejas, F., Nichols, E., Klavans, J., and Muresan, S. (2020). MorphAGram, evaluation and framework for unsupervised morphological segmentation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7112–7122, Marseille, France, May. European Language Resources Association.

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

Kazeminejad, G., Cowell, A., and Hulden, M. (2017). Creating lexical resources for polysynthetic languages—the case of Arapaho. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 10–18, Honolulu, March. Association for Computational Linguistics.

Lane, W. and Bird, S. (2020). Bootstrapping techniques for polysynthetic morphological analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6652–6661, Online, July. Association for Computational Linguistics.

Micher, J. (2017). Improving coverage of an Inuktitut morphological analyzer using a segmental recurrent neural network. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 101–106, Honolulu, March. Association for Computational Linguistics.

Schwartz, L., Tyers, F., Levin, L., Kirov, C., Littell, P., Lo, C.-k., Prud'hommeaux, E., Park, H. H., Steimel, K., Knowles, R., Micher, J., Strunk, L., Liu, H., Haley, C., Zhang, K. J., Jimmerson, R., Andriyanets, V., Muis, A. O., Otani, N., Park, J. H., and Zhang, Z. (2020). Neural polysynthetic language modelling.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

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

Snoek, C., Thunder, D., Lõo, K., Arppe, A., Lachler, J., Moshagen, S., and Trosterud, T. (2014). Modeling the noun morphology of Plains Cree. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Lan-*

guages, pages 34–42, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

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).

Virpioja, S., Turunen, V. T., Spiegler, S., Kohonen, O., and Kurimo, M. (2011). Empirical comparison of evaluation methods for unsupervised learning of morphology. *Traitement Automatique des Langues*, 52(2):45–90.

6. Language Resource References

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

Frantz, D. G. and Russell, N. J. (2017). *Blackfoot Dictionary of Stems, Roots, and Affixes*. University of Toronto Press, third edition.

Leavitt, R. (1996). *Passamaquoddy-Maliseet*. Languages of the world / Materials: Materials. Linde.