Evaluation of Morphological Segmentation Methods for Hupa

Nathaniel Parkes
Department of Linguistics
University of Florida
n.parkes@ufl.edu

Zoey Liu
Department of Linguistics
University of Florida
liu.ying@ufl.edu

Abstract

Building downstream NLP applications with tokenization systems built on morphological segmentation has been shown to be fruitful for certain morphologically-rich languages. Yet, indigenous and endangered languages, which tend to be highly polysynthetic and therefore potential beneficiaries of this approach, pose additional difficulties in their limited access to annotated data for morphological segmentation tasks. In this study, we develop morphological segmentation models for Hupa, a Dene/Athabaskan language critically endangered to North America. With a total of 595 word types, we seek to identify an optimal morphological segmentation model and illustrate how those tested perform under different levels of training data limitation. We propose a simple method that casts morphological segmentation as a sequence binary classification task. While this approach does not outperform the established practice of multi-class classification, it outperforms neural alternatives. This work is conducted under the intention to act as a starting point for future technological developments with Hupa looking to leverage its morphological qualities, which we hope can serve as a reflection for work with other indigenous languages being studied under similar constraints.

1 Introduction

The Hupa people of the Hoopa Valley Reservation in Humboldt county California are a federally recognized indigenous group within the United States with over 3,000 documented descendants (Encyclopaedia Britannica, 2024). Despite resistance to policies or attempts at cultural erasure imposed by the American Government, the Hupa tribe has shown signs of gradual increase in American influence, noted in reports dating back to the mid-20th century (Bushnell, 1968). Today, many aspects of their culture and tradition are upheld, but modern descendants are exhibiting a declining trend in language retention with English taking over as the primary language (Spence, 2021). Efforts are being made to revitalize this piece of their culture, but relevant language data is limited and the Hupa language, of the Dene/Athabaskan language family, is currently recognized under endangered status (Campbell and Grondona, 2008).

With the support of community members and linguists with advanced knowledge on the language, recent work has started to leverage computational techniques to facilitate documentation of Hupa and creation of pedagogical materials for language teaching. However, said research has only focused on automatic speech recognition (Venkateswaran and Liu, 2024). In this paper, we intend to contribute to such efforts, focusing specifically on morphological segmentation for Hupa. The goal of morphological segmentation is to automatically segment a word into its individual component morphemes (e.g., *lemons* $\rightarrow$ *lemon* + *s*).

Like many other native American languages, Hupa has a highly complex, yet productive, polysynthetic morphology (Goddard, 1902-1907). As a result, the process of segmenting words into their morphological components in Hupa is likewise a difficult process when completed manually by seasoned linguists. Building computational models to segment words into sub-words, or morphemes, can be advantageous for such morpheme-rich systems. Furthermore, this can have major implications in the automation of language documentation processes (see also Zevallos and Bel (2023)).

With that in mind, this study makes two contributions. First, we evaluate the performance of four different model alternatives for morphological segmentation for Hupa; we purposefully create experimental settings with varying degrees of data limitations in order to probe the robustness of these models when faced with severely resource-constrained contexts. Second, we propose a simple

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 188–193

augmentation to the sequence-tagging approach to morphological segmentation and show how it levels up to established neural techniques.

2 Related Work

The task of morphological segmentation has enjoyed popularity over the years for a number of reasons. First, morphological supervision has practical use in downstream NLP tasks such as dependency parsing (Seeker and Çetinoğlu, 2015) and language modeling (Blevins and Zettlemoyer, 2019). Morphological information has also been shown to be helpful for machine translation (Clifton and Sarkar, 2011; Mager et al., 2022) and automatic speech recognition (Afify et al., 2006), two tasks that are among some of the most useful for indigenous endangered speech communities (Zhang et al., 2021; Prud'hommeaux et al., 2021). In addition, morphological structures can be included in learning materials such as online dictionaries (Garrett, 2011).

Prior work has addressed morphological segmentation for low-resource morphologically complex languages, including cases such as Seneca (Liu et al., 2021) as well as Mexican indigenous languages (Kann et al., 2018). These studies largely focused on surface segmentation[1], where the concatenation of all the individual morphemes is the same as the initial surface word form (e.g., *lemons* → *lemon* + *s*). In this paper, we also concentrate on surface segmentation using orthographic representations of words in Hupa.

3 Experiments

3.1 Data and preprocessing

The data for this study consists of 595 word types (no duplicates), which were extracted from a set of nine unpublished Hupa texts drawn from archival manuscripts with handwritten transcriptions by Curtin (1888-1889), Goddard (1902-1907), Kroeber (1900-1906), and Woodward (1953), plus recorded and transcribed stories told by contemporary Hupa speaker Mrs. Verdena Parker and handwritten sources, both validated in consultation with Mrs. Parker. All transcriptions were rendered in the practical Hupa orthography originally developed in the 1980s by Victor Golla and the Hoopa Valley Tribe's language committee, which is featured in resources like the Hupa Online Dictionary and Texts Website[2] and the learner-oriented print dictionary on which it is based (Golla, 1996). The practical orthography uses conventions familiar to people who are already literate in English, and is accessible for a standard English keyboard, such as the use of the digraph *ch* for an alveopalatal affricate, *u* for a centralized schwa-like vowel in closed syllables, colon *:* for vowel length, and apostrophe *'* for glottalization of certain classes of consonants and glottal stops elsewhere. These orthographic representations were manually parsed into component morphemes. The complete dataset held an average of 3.10 morphemes per word, as well as an average of 4 characters per morpheme. Experiments were run using solely this practical orthographic transcription.

3.2 Dataset construction

To probe the impact of and the interaction between training data size and morphological segmentation methods, we create augmented datasets with varying training set sizes. We illustrate the dataset construction process with the following example.

Recall that the original dataset in orthographic representation for Hupa contains 595 unique items. We carry out the following procedures: (1) We first split this dataset evenly (roughly) into five folds; each time we select one fold as *the test set* and the concatenation of the other four folds as *the training data pool*. There are 595 / 5 = 119 items in each test set, thereby 476 items in each of the training data pools. (2) Based on the training data pool size, we decided on a range of training set sizes with mostly 100-item increment between each size: {100, 200, 300, 400, 476/training data pool size}. (3) With each training size, we randomly sample without replacement a training set of that size from a training data pool, 2 times, corresponding to two training sets of that size. (4) We repeated step (3) for each pair of training data pool + test set created from (1).

3.3 Model architectures

We study four model alternatives from two broad model classes: conditional random field (CRF) (Lafferty et al., 2001) and neural sequence-to-sequence (seq2seq) models.

CRF casts morphological segmentation as a sequence tagging task. Given a character w_t within a word w, where t indicates the index position of

[1] See Cotterell et al. (2016) for details on canonical segmentation.

[2] https://pages.uoregon.edu/jusp/dictionaries/hupa-lexicon.php

the character in the word, along with a curated feature set x_t that consists of n-grams of local (sub) strings, CRF gradually predicts the corresponding label y_t of the character using its feature set.

We curated the feature set for every character in each word as follows. We first appended each word with a start (<w>) and an end (</w>) symbol. The feature set for the character consists of the substring(s) occurring to the left and to the right side of the character up to a maximum length, δ. Consider the following Hupa word, $xotuq$, which consists of two morphemes: xo and tuq (them/people + between; together the word means "between them (people)"). If we were to set the value of δ to be 4 (which we did for model training), for the fourth character in the word, u, the sequence of substrings appearing to the left and to the right side of this character will be, respectively, {t, ot, xot, <w>xot} and {u, uq, uq</w>}. We concatenated these two sequences to be the full feature set of the fourth character u.

We implemented and compared two methods for character tagging here: multi-class classification, which is an approach applied before (Mager et al., 2022), and binary classification, inspired by Pranjić et al. (2024). With multi-class classification, for a character w_t at position t in word w, we assigned it one of six labels: `START` (for <w>); `END` (for </w>); `S` (for any single-character morpheme); and `B` (beginning); `M` (middle); or `E` (end) for characters in a multi-character morpheme. Based on the morpheme structure of the word $xotuq$, the segmentation labels are as follows:

<w>	x	o	t	u	q	</w>
START	B	E	B	M	E	END

In binary classification, said character w_t at position t in word w, if not set to `START` (for <w>) or `END` (for </w>), is assigned one of two labels: `B` (for any character bounded, or followed, by a morpheme boundary); and `U` (for characters unbounded, or not followed, by a morpheme boundary). Again, based on the morpheme structure of the word $xotuq$, labels are as follows:

<w>	x	o	t	u	q	</w>
START	U	B	U	U	U	END

We consider this form of classification as a simpler alternative to multi-class classification. If successful, breaking down the task of sequence tagging to a simple option of 0 or 1, bounded or unbounded, provides a more efficient data representation design that can possibly facilitate the model's training when faced with fewer resources.

We built first-order CRFs (Lafferty et al., 2001; Ruokolainen et al., 2013) for morphological segmentation. All models were implemented with the Python library `crfsuite`. This decision was motivated by two factors. First, prior work has demonstrated CRF to be superior to neural sequence-to-sequence models as well as different variants of unsupervised models such as Morfessor (Creutz and Lagus, 2002), when it comes to low-resource morphological segmentation for a variety of typologically diverse languages (Liu and Dorr, 2024; Liu and Prud'hommeaux, 2022; Cotterell et al., 2015). Second, CRF models, particularly those of lower orders (first-/second-order), are much faster and efficient to implement.

Our second model class is the neural-network models, specifically seq2seq. The models are expected to, given a word, produce an output of the equivalent word segmented by internal morpheme boundaries, indicated by the '!' delimiter below:

INPUT	x	o	t	u	q
OUTPUT	x	o	!	t	u	q

We made use of three seq2seq frameworks with the Python library `fairseq` (Ott et al., 2019), each under their default parameters: TRANSFORMER model (embedding size of 512, 6 encoder-decoder layers, 8 self-attention heads, and 2048 hidden units in the feed-forward layers); TRANSFORMER_TINY model, a less computationally demanding alternative contrary to the aforementioned (embedding dimension and feed-forward layer dimension both being 64; and a LSTM-based framework (embeddings of 512 dimensions and one hidden layer with 512 hidden units in both the encoder and the decoder).[3]

4 Results

We use $F1$ score as an evaluation metric for model performance. Table 1 shows the results of the CRF models for multi-class and binary classifications trained with differently sized training sets. Table 2 shows the results of the remaining three seq2seq models. Notably, the CRF models are most successful. Specifically, the multi-class classification CRFs outperform all other approaches/model architectures. While the binary classifier lags behind the multi-class alternative, it still performs notably better than any of the seq2seq models.

[3] https://fairseq.readthedocs.io/en/latest/models.html

Training Sample Size	*multiclass*	*binary*
100	70.07	62.08
200	76.18	68.96
300	78.40	70.13
400	80.27	71.16
Total	84.30	75.32

Table 1: Performance averages of the CRF-model architectures per Training Sample Size: multi-class classification and binary-classification; *Total* refers to the setting when all data from the training data pool is used for model training.

Sample Size	$Trans.$	$Tiny$	$LSTM$
100	7.41	15.46	9.63
200	13.96	28.64	15.49
300	20.56	39.58	25.37
400	29.54	46.15	34.60
Total	46.78	64.12	59.98

Table 2: Performance averages of the seq2seq-model architectures per Training Sample Size: TRANSFORMER ($Trans.$), TRANSFORMER_TINY ($Tiny$), and LSTM; *Total* refers to the setting when all data from the training data pool is used for model training.

Regarding the tendencies of the results between training sizes, we find that the CRF models showcase a gradual increase in performance capability as training set size increases. Despite CRF's sequence tagging strategy performing, comparatively, the most optimal in these low-resource environments, this trend demonstrates there is still a dependency on data set size to consider, with the dependency being stronger when training sizes are smaller (e.g., the largest $F1$ score increase occurs when training samples go from 100 to 200 word types).

The seq2seq models follow a similar trend, yet with much lower $F1$ score averages (Table 2). This is possibly due to the fact that neural-network models have much more complex training parameterization, which in turn can result in a reliance on much more extensive data resources (Wei and Ma, 2019). This conjecture is further supported by the results here that TRANSFORMER_TINY outperforms TRANSFORMER, with the former having a simpler architecture. The spread of $F1$ scores is also unique, with seq2seq models showing greater performance increase between larger training sets in comparison to what is observed with CRFs.

Learning that CRF models achieve the best performance in our experiments, we now ask: where do CRF models fall short? To address this question, we take a close look at the errors made by CRFs. Most remarkably, the CRF models struggle with words of 2 or more morpheme boundaries, especially those consisting of short, 1-3 character, morphemes. Around 66% of the time, the label-ers for both multi-class and binary classifications underestimate the number of morphemes in a word or simply predict words to be one single morpheme. Specifically, approximately 33% of all mistakes can be attributed to the later, in which CRFs fail to recognize the presence of any morpheme boundaries at all.

Another possible consideration of where the CRF models fall short is the lack of overlap between the training and the test sets. Almost none of the morphemes in the test sets can be found in the corresponding training data. With a lack of parallelism between model training and evaluation, this leaves ambiguity in certain morphological structural situations that segmentation models might fail to recognize. Yet, this challenge could be mended by data augmentation methods in the future.

5 Discussion & Future Directions

We attempt to provide evidence of the efficacy of various morphological segmentation models for Hupa and their level of robustness in response to different training set sizes. Our investigation identifies that CRF model performances shift in response to resource availability, yet they largely outperform neural alternatives in significantly low-resource settings. More notably, we also record a relatively successful CRF model using binary classification, again, outperforming all neural-network models. Despite not surpassing the multi-class classifier, the model averages are still relatively high and demonstrate a simple implementation which can be taken further in future work for Hupa and potentially other languages alike.

As mentioned prior, one caveat of model performance here is the recognition of words composed largely of short morphemes. To combat this issue, future work could consider experimenting with data in phonological representations in comparison to orthographic data. Phonological data formats may provide insight into phonetic environments for morpheme boundaries, providing suprasegmental details such as stress, tone, etc. Orthographic data may also falter as different sounds, varying in quality or length, are represented by the same symbol. Future experiments testing phonological datasets could help resolve ambiguity where morphological distinctions are created by phoneme variations that are not visible orthographically.

Another future direction for this study is to apply data augmentation methods to alleviate resource constraints. With a dataset of only 595 unique tokens, data augmentation could be implemented to strengthen validity of findings pertaining to the interaction of model performance and required data resources. In addition, while seq2seq models fell behind in this study, neural-networks may perform promisingly when trained under a larger artificially augmented dataset. We leave this for future work.

Finally, the findings reported in this paper and future avenues discussed are made with the purpose of continuously contributing to community-based efforts in language documentation. For the Hupa speech community, our plan for this line of work is to keep improving the performance of the morphological segmentation models, which will eventually be applied to automatically parse collected and digitized Hupa texts for use by the community. Additionally, we hope that our work will be helpful for other indigenous communities and academics engaged in similar efforts. To that end, we make all our code publicly available.[4]

Acknowledgments

We would like to express our gratitude to the Hupa indigenous community for their continuous support. We are grateful to Mrs. Verdena Parker and Dr. Justin Spence for their valuable efforts on Hupa language documentation over the years, and for giving us permission to work with the annotated data. Without their endorsement and expertise, this work would not have been possible. Lastly, we thank the reviewers for their thoughtful suggestions.

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*.

Terra Blevins and Luke Zettlemoyer. 2019. Better character language modeling through morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1606–1613, Florence, Italy. Association for Computational Linguistics.

John H. Bushnell. 1968. From American Indian to Indian American: The Changing Identity of the Hupa. *American Anthropologist*, 70(6):1108–1116.

Lyle Campbell and Verónica Grondona. 2008. Ethnologue: Languages of the world. *Language*, 84(3):636–641.

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, 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. 2016. 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 - Volume 6*, MPL '02, page 21–30, USA. Association for Computational Linguistics.

Jeremiah Curtin. 1888-1889. *Hupa vocabulary December 1888-January 1889*. National Anthropological Archives: NAA MS 2063.

Encyclopaedia Britannica. 2024. Hupa. https://www.britannica.com/topic/Hupa. Last updated: August 28, 2024.

Andrew Garrett. 2011. An online dictionary with texts and pedagogical tools: The Yurok language project at Berkeley. *International Journal of Lexicography*, 24(4):405–419.

Pliny Earle Goddard. 1902-1907. *Chilula field notes (Redwood Creek)*. American Philosophical Society Na20g.1.

Victor Golla. 1996. *Hupa Language Dictionary*. Hoopa, CA: Hoopa Valley Tribal Council.

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.

Alfred Kroeber. 1900-1906. *Untitled Hupa text*. Transcription in Kroeber's hand included in Goddard (1903-1906), notebook #4.

[4] https://github.com/ufcompling/hupa_morphseg

John Lafferty, Andrew McCallum, and Fernando C.N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

Zoey Liu and Bonnie Dorr. 2024. The effect of data partitioning strategy on model generalizability: A case study of morphological segmentation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2851–2864, Mexico City, Mexico. Association for Computational Linguistics.

Zoey Liu, Robert Jimerson, and Emily Prud'hommeaux. 2021. Morphological segmentation for Seneca. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 90–101, Online. Association for Computational Linguistics.

Zoey Liu and Emily Prud'hommeaux. 2022. Data-driven Model Generalizability in Crosslinguistic Low-resource Morphological Segmentation. *Transactions of the Association for Computational Linguistics*, 10:393–413.

Manuel Mager, Arturo Oncevay, Elisabeth Mager, Katharina Kann, and Thang Vu. 2022. BPE vs. morphological segmentation: A case study on machine translation of four polysynthetic languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 961–971, Dublin, Ireland. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Marko Pranjić, Marko Robnik-Šikonja, and Senja Pollak. 2024. LLMSegm: Surface-level morphological segmentation using large language model. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10665–10674, Torino, Italia. ELRA and ICCL.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic speech recognition for supporting endangered language documentation. *Language Documentation and Conservation*, 15:491–513.

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.

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.

Justin Spence. 2021. A Corpus Too Small: Uses of Text Data in a Hupa-English Bilingual Dictionary. *International Journal of Lexicography*, 34(4):413–436.

Nitin Venkateswaran and Zoey Liu. 2024. Looking within the self: Investigating the impact of data augmentation with self-training on automatic speech recognition for Hupa. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 58–66, St. Julians, Malta. Association for Computational Linguistics.

Colin Wei and Tengyu Ma. 2019. Data-dependent sample complexity of deep neural networks via Lipschitz augmentation. *CoRR*, abs/1905.03684.

Mary F. Woodward. 1953. *Survey of California and Other Indian Languages*. University of California Berkeley, Woodward.002.

Rodolfo Zevallos and Nuria Bel. 2023. Hints on the data for language modeling of synthetic languages with transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12508–12522, Toronto, Canada. Association for Computational Linguistics.

Shiyue Zhang, Benjamin Frey, and Mohit Bansal. 2021. ChrEnTranslate: Cherokee-English machine translation demo with quality estimation and corrective feedback. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 272–279, Online. Association for Computational Linguistics.