Recovering Missing Characters in Old Hawaiian Writing

Brendan Shillingford[*1,2] and **Oiwi Parker Jones**[*1]
University of Oxford[1], DeepMind[2]
brendan.shillingford@cs.ox.ac.uk,
oiwi.parkerjones@wolfson.ox.ac.uk

Abstract

In contrast to the older writing system of the 19th century, modern Hawaiian orthography employs characters for long vowels and glottal stops. These extra characters account for about one-third of the phonemes in Hawaiian, so including them makes a big difference to reading comprehension and pronunciation. However, transliterating between older and newer texts is a laborious task when performed manually. We introduce two related methods to help solve this transliteration problem automatically. One approach is implemented, end-to-end, using finite state transducers (FSTs). The other is a hybrid deep learning approach, which approximately composes an FST with a recurrent neural network language model.

1 Introduction

From 1834 to 1948, more than 125,000 newspaper pages were published in the Hawaiian language (Nogelmeier, 2010). Yet by 1981, many expected this once flourishing language to die (Benton, 1981). Hawaiian has since defied expectations and experienced the beginnings of a remarkable recovery (Warner, 2001; Wilson and Kamanā, 2001). However much of the literary inheritance that is contained in the newspapers has become difficult for modern Hawaiians to read, since the newspapers were written in an orthography that failed to represent about one-third of the language's phonemes. This orthography, which we will refer to as the *missionary orthography*, excluded Hawaiian phonemes that did not have equivalents in American English (see Schütz, 1994), including Hawaiian's long vowels /iː eː aː oː uː/ and glottal stop /ʔ/. By contrast, the *modern Hawaiian orthography*, an innovation of Pukui and Elbert's Hawaiian dictionary (Pukui and Elbert, 1957), presents a nearly perfect, one-to-one mapping between graphemes and phonemes (see Appendix A.1). The process of manual transliteration from missionary to modern Hawaiian orthography is extremely labor intensive. Yet the cultural benefits are so great that hundreds of pages of newspaper-serials have already been transliterated by hand, such as Nogelmeier's new edition of the epic tale of *Hi'iakaikapoliopele*, the volcano goddess's sister (Ho'oulumāhiehie, 2007). Critically important as such efforts are to the continued revitalization of this endangered language, they are still only an introduction to the material that could be translated for a modern Hawaiian audience.

In this paper, we propose to automate, or semi-automate, the transliteration of old Hawaiian texts into the modern orthography. Following a brief review of related work (Section 2), we begin by describing a dataset of modern Hawaiian (Section 3). In Section 4, we present two methods for recovering missing graphemes (and hence phonemes) from the missionary orthography. The first composes a series of weighted FSTs; the second approximately composes a FST with a recurrent neural network language model (RNNLM) using a beam search procedure. Both approaches require only modern Hawaiian texts for training, which are much more plentiful than parallel corpora. Section 5 reports the results of our transliteration experiments using a simulated parallel corpus, as well as two 19th century newspaper articles for which we also have modern Hawaiian transcriptions. Being based on FSTs, both approaches are modular and extensible. We observe useful and promising results for both of our methods, with the best results obtained by the hybrid FST-RNNLM. These results showcase the strength of combining established hand-engineering methods with deep learning in a smaller data regime, with practical applications for an endangered language.

[*]Authors contributed equally.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4929–4934
Brussels, Belgium, October 31 - November 4, 2018.

2 Related work

Many of the themes that we address relate to existing literature. For example, Hajič et al. (2000) and Scannell (2014) have written on machine translation (MT) for closely related languages and on multilingual text normalization. Though language-relatedness makes MT easier (Kolovratník et al., 2010), state-of-the-art techniques such as neural machine translation (NMT) have not performed well for languages with little data (Östling and Tiedemann, 2017). So while the Hawaiian transliteration problem could be cast as an instance of MT or of NMT, we chose to sidestep the scarcity of parallel data by not considering such approaches.

Hybrid approaches that combine expert knowledge for well-understood structures with deep learning for data-plentiful subproblems offer rich opportunities for data-efficient modelling. Prior work has combined FSTs with RNNs, although not using the approximate FST-to-RNN composition algorithm that we introduce here (in Appendix A.4). For example, Sproat and Jaitly (2016) used an FST to restrict the search space when decoding from an RNN and Rastogi et al. (2016) incorporated RNN information into an FST.

3 Data

3.1 Missionary & modern orthography

The primary difference between the missionary and modern Hawaiian orthographies is that the missionary orthography does not encode long vowels or the glottal stop (see Appendix A.1). For example, the following Hawaiian phrases were recorded by a 19th-century German traveller in the missionary orthography: *Ua oia au*, *E ue ae oe ia Ii*, *E ao ae oe ia ia* (Chamisso, 1837, p. 7). In the modern orthography these become: *Ua ʻō ʻia au* 'I am speared', *E uē aʻe ʻoe iā ʻĪʻī* 'You must weep for 'Īʻī (a person)', and *E aʻo aʻe ʻoe iā ia* 'You teach him' (Elbert and Pukui, 1979, p. 3).

We can convert text in the modern Hawaiian orthography *backward* chronologically to an approximate missionary orthography by mapping each glottal stop ⟨ʻ⟩ to the empty string ε, and each long vowel, e.g. ⟨ā ē ī ō ū⟩, to its corresponding short vowel, ⟨a e i o u⟩. As a first approximation, we may treat mappings from the modern-to-missionary orthographies as unambiguously many-to-one; thus there is information loss. We will return to secondary differences between the orthographies in Section 6. To illustrate, the following four words in the modern orthography all map to the same missionary string *aa*: *aʻa* (root), *ʻaʻa* (brave), *ʻaʻā* (crumbly lava rock), and *ʻāʻā* (stutter).

Source	Chars	Words
Ulukau(160 texts)	6,518,451	1,334,451
Hiʻiakaikapoliopele	1,272,935	259,947
Wikipedia	577,794	10,221
Total	8,369,180	1,604,619

Figure 1: Modern data sources and their sizes.

The *forward* mapping from missionary-to-modern orthographies is one-to-many. Thus the missionary string *aa* could map to *aʻa*, *ʻaʻa*, *ʻaʻā*, or *ʻāʻā*. The *transliteration problem* we address here seeks to discover how we can use context to recover the information not present in the missionary orthography that modern Hawaiian orthography retains.

3.2 Data sources

We draw on three sources for modern Hawaiian text: the main text of *Hiʻiakaikapoliopele* (Hoʻoulumāhiehie, 2007), 160 short texts from *Ulukau: The Hawaiian Electronic Library*, and the full Hawaiian Wikipedia (see Figure 1).[1]

For evaluation, we simulate a missionary-era version of the modern texts using the backward mapping described above. In addition, we evaluated our models on a couple of 19th century newspaper samples for which we have parallel missionary-era and modern text. Both simulated and real parallel corpora will be described in Section 5.

4 Models

We can frame the task of transliterating from missionary-to-modern Hawaiian orthographies as a sequence transduction problem. Many deep learning approaches (e.g. Sutskever et al., 2014; Graves, 2012) are not easily applicable to this task since we do not have a sufficiently large dataset of parallel texts. Instead, we focus on approaches that mix hand-designed finite state transducers with trained language models, including deep learning approaches like RNNLMs (Mikolov et al., 2010).

[1] *Ulukau: The Hawaiian Electronic Library*: http://ulukau.org/, Hawaiian Wikipedia: https://haw.wikipedia.org/. Both accessed 19 May 2018.

4.1 Finite state transducers

Our initial approach represents the mapping from missionary to modern orthography using a composition of (weighted) FSTs. For a thorough review of FSTs, see Mohri (1997).

First, we construct a finite state acceptor, I, from the input text. Here we construct a trivial chain-shaped acceptor that accepts only the input text. Each symbol in the input text is represented by a state which emits this symbol on a single transition that moves to the next state. The transition emitting the final symbol in the string leads to the sole accepting state.

Second, we construct a missionary-to-modern orthography conversion FST which we call C, which models potential orthography changes that can occur when transliterating from the missionary to modern Hawaiian orthography. For example, two non-deterministic transitions introduce an optional long-vowel map $(\text{a} : \bar{\text{a}})$ and $(\text{a} : \text{a})$. Another transition inserts glottal stops: $(\epsilon : \text{‘})$. By capturing the orthographic changes we know to occur, the composition $I \circ C$ produces a large set of candidates to be narrowed using the language model.

Third, we use the modern Hawaiian text from Section 3.2 to construct and evaluate a number of character-level n-gram language models, of various combinations of order and Katz backoff and Kneser-Ney (KN) smoothing (Katz, 1987; Kneser and Ney, 1995); see Appendix A.5 for details. N-gram language models can be expressed as weighted FSTs. We denote the n-gram or weighted FST language model as G. Character-level models are used as we wanted to generalize to out-of-vocabulary words, which we expected to occur frequently in a small corpus like the one we have for Hawaiian.

Finally, we use this model to infer modern orthography given a piece of text in missionary orthography as input, then compose the FSTs to form the *search graph* FST: $S = I \circ C \circ G$. The minimum cost path through S gives the predicted modern orthography. Of these n-gram-based approached, we found the Kneser-Ney-based models to perform best; these approaches are called FST-C-NGRAM-KN and FST-C_{wb}-NGRAM-KN.

We circumvent the lack of a large, non-simulated parallel corpus by training the language model exclusively on text in the modern Hawaiian orthography. In turn, the orthographic transliteration FST C produces candidates which are disambiguated by the language model. The result is finally evaluated against the ground-truth modern text.

Although the orthographic transliteration model is an approximation, and thus not exhaustive, it embodies an explicit and interpretable representation that can be easily extended independently of the rest of the model. To illustrate how the approach can be extended, we constructed a variant C_{wb} (where wb stands for word boundary). C_{wb} optionally inserts a space after each vowel using an additional arc that maps $(\epsilon : \textit{space})$, as diagrammed in Appendix A.2. This variant is able to model some changes in Hawaiian's word-boundary conventions (Wilson, 1976), such as *alaila* becoming *a laila* which demarcates the preposition *a* 'until' and noun *laila* 'then'. We employ this variant to predict modern equivalents from 19th century newspaper samples in Section 5. Pseudocode summarizing this method is shown in Appendix A.3. Example predictions can be found in Appendix A.6.

4.2 FSTs with LSTM language models

As an alternative approach, we combined the FST C in the previous section with an RNNLM (Mikolov et al., 2010). RNNLMs often generalize better than n-gram models.

An RNN is a neural network that models temporal or sequential data, by iterating a function mapping a state and input to a new state and output. These can be stacked to form a deep RNN. For language modelling, each step of the final RNN layer models a word or character sequence via $p(w_1, \ldots, w_n) = \prod_{i=1}^{n} p(w_i | w_{1:i-1})$ and can be trained by maximum likelihood. Recent language modeling work has typically used the long short-term memory (LSTM) unit (Hochreiter and Schmidhuber, 1997) for its favorable gradient propagation properties. All RNNs in this paper are LSTMs.

Our goal is to replace the n-gram language model in the end-to-end FST approach with an RNNLM. While the minimum cost path through an FST can be computed exactly as done in the previous section, it is not straightforward to compose the relation defined by an FST with an arbitrary one like that defined by an RNNLM. A minimum cost path through the composition of the FST and the RNNLM can be defined as a path (i.e. label sequence) that minimizes the sum of the FST path cost and the RNNLM cost.

We can approximately find a minimum cost path of the composition of the two models by a breadth-first search over the FST graph, or using a beam search, as follows. At any particular iteration, consider a single beam element. The beam element holds the current FST and RNN states, and the path taken through the FST so far. We follow each possible arc from the current FST state, each producing a new child beam element, and feed the output symbol into the RNN (unless it is ϵ). There may be duplicate beam elements due the nondeterminicity of the FST; in this case, the lower cost edge wins. We sort by the sum of the FST and RNN costs, keep the lowest-cost K, and then proceed to the next iteration. If a beam element is on an accepting state of the FST, it is kept as-is between iterations. Detailed pseudocode is provided in Appendix A.4.

In the following we will refer to the hybrid models as FST-RNNLM—or as FST-RNNLM-C and FST-RNNLM-C_{wb} if we want to distinguish between which FST we used. Similarly, the FST-only models will be referred to as FST-C and FST-C_{wb}, with suffixes denoting what kind of n-gram and smoothing were used. For example, FST-C-7GRAM-KN denotes a FST-only model with an 7-gram language model and Kneser-Ney smoothing. Details of the language models trained can be found in Appendix A.5.

5 Results

Evaluation. Since we were unable to find a sufficiently large corpus of parallel texts in the missionary and modern Hawaiian orthographies, we instead used a corpus of modern Hawaiian texts (*ground-truth*) as summarized in Section 3.2 and Figure 1. Note that training the n-gram and RNN language models required only this modern corpus.

To evaluate the accuracy of our approaches, we derived a synthetic parallel corpus from these modern Hawaiian texts. We also used a small but real parallel corpus, based on two 19th century newspaper texts and their hand-edited modern equivalents.

Simulated parallel corpus. To produce a simulated parallel corpus (*input-missionary*), we systematically reduced the orthography in the modern texts using the backward mapping described in Section 3.1. We then applied the two approaches described in Section 4, with the aim of recovering the information lost.

We evaluated the predicted modern text (*predictions*) by computing

$$\text{CERR} = \frac{d(\text{prediction}, \text{ground-truth})}{d(\text{input-missionary}, \text{ground-truth})},$$

where d denotes character-level edit distance. This is a modification of character error rate, normalized by the distance of the input and target rather than by the length of the target. We note that CERR may be high even when the predictions are very accurate as $d(\text{input-missionary}, \text{ground-truth})$ is small when the text is similar in both orthographies.

Table 1 reports the results of the approaches we described in Section 4. Out of the Kneser-Ney n-gram models, we found that the FST-C-9GRAM-KN and the version modelling word boundaries (FST-C_{wb}-9GRAM-KN) to perform best on the synthetic parallel corpus and newspapers, respectively. C_{wb} was not applied to the synthetic parallel corpus as we did not model word splitting. The hybrid models (FST-RNNLM) outperformed all FST-only approaches.

Real parallel corpus (newspaper texts). Not content to evaluate the model on simulated missionary orthography, we also evaluated it on two newspaper texts, using selections originally published in 1867 and 1894 for which we had 19th century and manually-edited modern equivalents. The newspaper selections discuss *Kahahana*, one of the last kings of O‘ahu (Kamakau and Perreira, 2002), and *Uluhaimalama*, a garden party and secret political gathering, held after the deposition of Hawai‘i’s last queen (Pukui et al., 2006). Unlike the synthetic missionary corpus evaluation where we did not model word splitting, we found that replacing C with C_{wb} on the newspaper texts significantly improved the output, especially on the FST-RNNLM model. Thus, we found the word-splitting hybrid model (FST-RNNLM-C_{wb}) to be the best performing model overall (Table 1).

6 Conclusions and future work

With this paper we introduced a new transliteration problem to the field, that of mapping between old and new Hawaiian orthographies—where the modern Hawaiian orthography represents linguistic information that is missing from older missionary-era texts. One difficulty of this problem is that there is a limited amount of Hawaiian data, making data-hungry solutions like end-to-end deep learning

Transliteration model	LM perplexity Valid.	Test	Transliteration performance (%CERR) Corpus	Newspaper 1	Newspaper 2
FST-(C/C_{wb})-7GRAM-KN	3.07	3.13	27.3%	50.1% / 38.7%	52.0% / 47.5%
FST-(C/C_{wb})-9GRAM-KN	2.95	3.02	26.6%	50.7% / 39.3%	52.5% / 47.2%
FST-(C/C_{wb})-11GRAM-KN	2.94	3.02	27.8%	53.9% / 41.3%	54.1% / 48.7%
FST-RNNLM-(C/C_{wb})	**2.65**	**2.69**	**16.3%**	47.2% / **34.3%**	49.8% / **41.2%**

Table 1: Performance (%CERR). Slash-separated pairs denote FSTs incapable/capable of inserting word boundaries, respectively; see Section 4. The -KN suffix denotes Kneser-Ney smoothing. The data from Section 3.2 is used for evaluating the modern-orthography language model perplexity, and "Corpus" evaluates test-set transliteration performance from the synthetic missionary text back to the original modern text.

Input	Ua lawe ola ia o Keawehano imua o Kahekili, a ua hai aku o Kapohu...
Prediction	Ua lawe ola ʻia ʻo Keawehano i mua o Kahekili, a ua haʻi aku ʻo Kapohu...
Ground-truth	Ua lawe ola ʻia ʻo Keawehano i mua o Kahekili, a ua haʻi aku ʻo Kapohū...

Figure 2: An example of (missionary input, predicted modern text, ground-truth), from each newspaper. Note the correctly split word in the second example. Incorrect characters, which are quite rare, are shown as red and underlined. More sample predictions can be found in Appendix A.6.

unlikely to work. To solve the transliteration problem, we therefore proposed two models: the first was implemented end-to-end using weighted FSTs; the second was a hybrid deep learning approach that combined an FST and an RNNLM. Both models gave promising results, but the hybrid approach, which allowed us to use a more powerful recurrent neural network-based language model despite our dataset's small size, performed best. Factoring a problem like ours into one part that can be modelled exactly using expert domain knowledge and into another part that can be learned directly from data using deep learning is not novel; however it is a promising research direction for data-efficient modelling. To our knowledge, this paper is the first to describe a procedure to compose an FST with an RNN by approximately performing beam search over the FST.

While the role of the RNNLM part of the hybrid approach may be obvious, the FST component plays an important role too. For example, the hand-designed FST component can be replaced without needing to retrain the RNNLM. We tried to showcase this modularity by constructing two FSTs which we referred to as C and C_{wb}, where only the latter allowed the insertion of spaces. Future work could extend the FST to model orthographic changes suggested by an error analysis of the current model's predictions (see Appendix A.6). These errors motivate new mappings for consonant substitutions like (r : l) and (s : k) observed in loanword adaptations (e.g. *rose* ⇒ *loke*). The error analysis also motivates mappings to delete spaces ($\text{␣} : \epsilon$) and to handle contractions, like *na'lii* ⇒ *nā ali'i*. We could further incorporate linguistic knowledge of Hawaiian into the FST, which tells us, for example, that a consonant is typically followed by a vowel (Parker Jones, 2010). Additional improvements to the hybrid model might be obtained by increasing the amount of modern Hawaiian text used to train the RNNLM. One way to do this would be to accelerate the rate at which missionary-era Hawaiian texts are modernized. To this end, we hope that the present models will be used within the Hawaiian community to semi-automate, and thereby accelerate, the modernization of old Hawaiian texts.

Acknowledgments

We are grateful to M. Puakea Nogelmeier for providing an electronic copy of *Hi'iakaikapoliopele* (Ho'oulumāhiehie, 2007).

References

Richard A Benton. 1981. *The flight of the Amokura: Oceanic languages and formal education in the South Pacific*. New Zealand Council for Educational Research, Wellington.

Adelbert von Chamisso. 1837. *Über die Hawaiische Sprache, Vorgelegt der Königlichen Academie der Wissenschaften zu Berlin am 12, Januar, 1837*. Weidmann, Leipzig.

Samuel H. Elbert and Mary Kawena Pukui. 1979. *Hawaiian Grammar*. University of Hawai‘i Press, Honolulu.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.

Jan Hajič, Jan Hric, and Vladislav Kuboň. 2000. Machine translation of very close languages. In *ANLC '00 Proceedings of the sixth conference on Applied natural language processing*, pages 7–12.

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

Ho‘oulumāhiehie. 2007. *Ka Mo‘olelo o Hi‘iakaikapoliopele*. Awaiaulu Press, Honolulu. Edited by M. Puakea Nogelmeier.

Samuel Manaiakalani Kamakau and Hiapo Perreira. 2002. Ka mo‘olelo o Kahahana, māhele 1. *Ka Ho‘oilina*, 1(1):102–121.

Slava Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE transactions on acoustics, speech, and signal processing*, 35(3):400–401.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 181–184. IEEE.

David Kolovratník, Natalia Klyueva, and Ondřej Bojar. 2010. Statistical machine translation between related and unrelated languages. In *Proceedings of the Conference on Theory and Practice of Information Technologies (ITAT-09)*, pages 31–36.

Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Černocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proceedings of Interspeech*, pages 1045–1048.

Mehryar Mohri. 1997. Finite-state transducers in language and speech processing. *Computational linguistics*, 23(2):269–311.

M. Puakea Nogelmeier. 2010. *Mai Pa‘a i ka Leo: Historical Voice in Hawaiian Primary Materials: Looking Forward and Listening Back*. Bishop Museum Press, Honolulu.

Robert Östling and Jörg Tiedemann. 2017. Neural machine translation for low-resource languages. *Computing Research Repository*, arXiv:1708.05729. Version 1.

Oiwi Parker Jones. 2010. *A computational phonology and morphology of Hawaiian*. Ph.D. thesis, University of Oxford.

Oiwi Parker Jones. 2018. Illustrations of the IPA: Hawaiian. *Journal of the International Phonetic Association*, 48:103–115.

Mary Kawena Pukui and Samuel H. Elbert. 1957. *Hawaiian-English Dictionary*. University of Hawai‘i Press, Honolulu.

Mary Kawena Pukui, Holo Ho‘opai, Oiwi Parker Jones, and Keao NeSmith. 2006. No ka mahi‘ai ‘ana, māhele 6. *Ka Ho‘oilina*, 5(1):2–23.

Pushpendre Rastogi, Ryan Cotterell, and Jason Eisner. 2016. Weighted finite-state transductions with neural context. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 623–633.

Kevin Scannell. 2014. Statistical models for text normalization and machine translation. In *Proceedings of the First Celtic Language Technology Workshop*, pages 33–40.

Albert J. Schütz. 1994. *The Voices of Eden: A history of Hawaiian language studies*. University of Hawai‘i Press, Honolulu.

Richard Sproat and Navdeep Jaitly. 2016. RNN approaches to text normalization: A challenge. *Computing Research Repository*, arXiv:1611.00068. Version 2.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Sam L. No‘eau Warner. 2001. The movement to revitalize Hawaiian language and culture. In Leanne Hinton and Kenneth Hale, editors, *The Green Book of Language Revitalization in Practice*, pages –144. Academic Press, San Diego, CA.

William H. Wilson. 1976. Standardized Hawaiian orthography. Manuscript, University of Hawai‘i.

William H. Wilson and Kauanoe Kamanā. 2001. “Mai loko mai o ka ‘i‘ini: Proceeding from a dream”: The ‘Aha Pūnana Leo connection in Hawaiian language revitalization. In Leanne Hinton and Kenneth Hale, editors, *The Green Book of Language Revitalization in Practice*, pages 147–176. Academic Press, San Diego, CA.