Automatic Lexicon Construction for Endangered Languages

Riya Patel
Carnegie Mellon University
riyap@cs.cmu.edu

Swetha Gangu
Carnegie Mellon University
sgangu@andrew.cmu.edu

Nathaniel Robinson
Carnegie Mellon University
nrrobins@cs.cmu.edu

Abstract

For endangered languages, data is scarce, and lexicons are often not available, which makes it difficult to develop automatic transcription systems. WFST-based methods leverage phoneme lattices and translations of minority language speech into higher-resourced languages to develop a transcription pipeline. We take advantage of recent developments in language-universal automatic speech recognition (ASR) to adapt these methods to new endangered languages. In particular, we focus on Tulu and gear our experiments towards effective Tulu ASR to improve education and language revitalization efforts. This work has profound potential for Tulu's 1.7 million speakers. We will release our code in a public repository[1].

1 Introduction and Problem Statement

Language documentation is necessary for preserving endangered languages with little to no resources. Speech data for these languages is often relatively easy to collect, but transcribing this data to document lexicons, morphology, glosses, and other linguistic features is prohibitively slow. Michaud *et al.* (2014) estimated that it takes a trained linguist roughly 1 hour to transcribe 1 minute of speech. Language documentation and revitalization efforts can be aided greatly by automatic transcription systems.

We build on the work by Adams *et al.* (2016b) to improve automatic phoneme transcription using bilingual data to learn a lexicon and translation model from phoneme lattices and target translations. In this approach, a Bayesian nonparametric model represents the joint distribution between source acoustic features, phonemes, and latent source words given target words. We update this pipleline by using universal phone recognition models developed by Li *et al.* (2020). Rather than mapping from speech data to language specific phonemes, we map to language universal phones and fine-tune weighted finite-state transducer (WFST) allophone graphs [Yan *et al.*, 2021] to map to from these to language-specific phonemes. This approach is generalizable and allows for rapid adaption to endangered languages. It will also make for robust models by leveraging advantages of shared parameters across languages. Its application will facilitate lexicon production for endangered languages for which education materials are scarce, providing them an avenue towards revitalization.

1.1 The Case of Tulu

On such language is Tulu. One of the UNESCO "Atlas of the World's Languages in Danger,"[2] Tulu is spoken in the states of Karnataka and Kerala in Southern India and has roughly 1.7 million speakers according to a 2001 census. UNESCO considers the degree of endangerment for Tulu to be "vulnerable", meaning that most children speak the language but usage is typically restricted to home. Most Tulu speakers, similar to speakers of other endangered languages, are bilingual and use the lingua franca of the regions (Kannada and Malyalam) outside of the home. As an endangered language, Tulu is not taught in academic settings and its written script Tigalari is no longer in use. As a result, there is a high risk of Tulu becoming extinct in the coming decades. Improved lexical documentation of Tulu and translation models to the lingua franca can be useful as a learning resource to strengthen knowledge of Tulu in regions where it was originally spoken.

Language extinction represents a massive and irretrievable loss of culture, knowledge, and history. This work aims to achieve automatic lexicon construction for endangered languages such as Tulu

[1] https://github.com/n8rob/latticetulu

[2] https://en.unesco.org/news/new-updates-available-interactive-atlas-world-s-languages-danger

from speech data aligned with translations to a more high-resource language, in order to mitigate that risk. Tulu is a particularly interesting language to use in this case because there are already strong connections from the language to higher-resource languages such as Kannada and English (a language that most Tulu films are translated into).

2 Literature Survey

We surveyed a number or related works pertaining to our central focus of phoneme recognition via WFST models. In addition to this primary focus, we review related work on Tulu language processing and methods to leverage well-trained language-universal ASR techonologies for our purposes.

2.1 ASR Via Translation Data and Phoneme Lattices

Adams *et al.* (2016b) developed an automatic transcription pipeline that learns transcription phonemes, given audio data and transcribed translations. Their Bayesian non-parametric model expressed as a weighted finite-state transducer (WFST) predicts phonemes and latent textual transcriptions. Their modular approach models three distributions and maximizes their product:

$$\hat{\phi} = \arg\max_{\phi,f} P(x|\phi)P(\phi|f)P(f|e) \qquad (1)$$

where ϕ is transcription phonemes, x is reference audio, f is latent textual transcription, and e is translation text. $\hat{\phi}$ is the set of phonemes predicted to match x. This method of phoneme prediction using translation data improved word error rate (WER) over ablations. The authors applied the framework for Japanese and German ASR but did not apply it to any endangered languages. And because they do not apply the approach with a universal phone recognizer, their method is not truly generalizable. To address this limitation, we leverage recent work in language-universal phone-based ASR to apply this strategy to endangered languages.

This method builds on prior work completed on phoneme translations modeling [Duong *et al.*, 2016], computer-aided speech translation [Pelemans *et al.*, 2015], translation modeling from automatically transcribed speech [Paulik and Waibel, 2013], Bayesian word alignment [Mermer and Saraclar, 2011; Zezhong *et al.*, 2013], and language model learning from lattices [Neubig *et al.*, 2012].

There has also been extensive work on utilizing translation models to improve ASR performance [Vidal *et al.*, 2006]. However, this prior work does not use phoneme lattices.

Other previous work with lattices includes research to learn translation models from ASR word path lattices via WFSTs [Adams *et al.*, 2016a]. In this related setup, models predict f by approximating

$$\hat{f} = \arg\max_{f} P(e|f)P(x|f)P(f) \qquad (2)$$

but probability calculation and parameter learning are performed similarly to the work of Adams *et al.* (2016b). Word lattices are used instead of phoneme lattices.

Our work will extend this vein of research by incorporating newer methods for phone recognition and phone-to-phoneme mapping.

2.2 Universal Phone Recognition

Recent universal phone recognition methods allow us to leverage pre-trained models for low-resource applications, exploiting transfer learning from higher-resourced languages. The universal phone recognition model with multilingual allophone system by Li *et al.* (2020), or Allosaurus, improved performance by 17% phoneme error rate over the baseline conditions for unseen low-resourced testing languages. Rather than mapping from speech data to language-specific phonemes using limited data from the low-resource language, the universal phoneme recognizer is used to map language-universal phones to language-specific phonemes.

2.3 Adaption of Phone Recognition to Phonemes

Yan *et al.* (2021) developed a new method called AlloGraph to map language non-specific phones to language-specific phones via WFST and is designed to be incorporated with language-universal ASR systems like Allosaurus (see section 2.2). They method improves performance over using matrices of phone-to-phoneme encodings, and the authors demonstrated rapid adaptability of the WFST AlloGraphs to unseen languages. We will apply a number of strategies to apply this same approach to endangered languages like Tulu. (See section .)

2.4 Tulu Language Models

There has been limited research done on Tulu language, especially since it is endangered and has very limited resources. Antony *et al.* (2012) completed previous work in constructing a morphological analyzer and generator (MAG) for Tulu using AT&T OpenFST[3]. Their work focused on using a rules-based approach since there does not exist a large aligned, well-generated corpus to perform corpus-based methods of MAG. The system utilized lexicon, morphotactics, and orthographic rules and was able to develop written rules for all noun and verb forms via FST. This rules-based MAG can analyze and generate a large majority of nouns and verbs used in daily conversation, including those with continuous and negative words. The MAG proposed in this work may be used to help in machine translation between Tulu to a more documented language like Kannada.

Shivakumar (2010) worked on expanding IndoWordNet to create a WordNet for the Tulu language. Since there were no previous lexical studies in Tulu, they used Hindi WordNet as a baseline. This was necessary since there are very few linguists working on Tulu; resources for Tulu synonymy, antonymy, hypernymy, etc. are not well documented. Additionally, since the Tulu script is rarely used, Shivakumar used Kannada script for Tulu transcription. Resources were limited to a few dictionaries mapping Tulu to English, Tulu to Kannada, and English to Kannada, and two corpora of Tulu folk songs and tales with English translation. As a result, when creating the WordNet, Shivakumar gave equivalents in both Tulu and Kannada to allow for cross-checking when recognizing complex lexical units and compounds.

Both of these works rely on transcribed Tulu and suffered from limited availability of transcriptions. Our work on Tulu ASR can open up immensely greater possibilities for transcribed Tulu training data in order to improve applications such as MAG and lexicon documentation.

3 Baseline Experiment

3.1 Experiment Structure

As a baseline experiment, we implemented Adams *et al.*'s (2016b) pipeline for phoneme recognition from speech, described in Section 2. We used the C++ implementation available on GitHub[4], and we used the OpenFST library[5].

Because the data set used for phoneme prediction in this work was not made publicly available, we used the Fisher data set containing Spanish word lattices and their corresponding Spanish sentences and English translations. This is the same data set used for word prediction by Adams *et al.* (2016a), and as such we applied the phoneme recognition method to word lattices.

The Spanish lattices in the corpus were given in Joshua lattice format and thus had to be converted to openFST lattice format to be made compatible with the pipeline. Additionally, the corpus had numerous alignment issues where the Spanish lattices did not match up with the English translations. This rendered a large part of it to be useless. We manually aligned the dev set and part of the test set for use in experimentation. The aligned pairs we submitted to the pipeline were split with 90 of the data being used for train and 10% of the data being used for test. The model was trained for only 5 epochs.

Aligned pairs	*WER*
100	53.4%
300	57.0%
700	47.9%
1200	49.1%
1700	46.3%
2500	61.9%

Table 1: Word error rate for ASR using different amounts of training data. Aligned pairs are Spanish audio mapped to English transcriptions, with word lattices given. Spanish transcriptions are predicted. For all processes we used a 90-10 train-eval split on total data and trained for 5 epochs.

3.2 Results and Error Analysis

The metric used to evaluate the results was WER. The Spanish sentences generated from the Spanish word lattices and English translations were compared to the gold standard Spanish translations. In order to achieve meaningful WERs, both the Spanish translations and the output of the model had to be normalized by using Unidecode to remove accents.

Results for our experiments are in Table 1. Passing more data into the pipeline seemed to generally bring down the WER with the exception of some outliers. Perhaps running for more epochs

[3]https://www.openfst.org/twiki/bin/view/Contrib/FsmLibrary
[4]https://github.com/oadams/latticetm
[5]https://www.openfst.org/

or with larger datasets would reveal a more clear relationship between training epochs, dataset size, and word error rate. This model seems to perform fairly well in low-resource conditions. Smaller sentences seemed to be more accurately constructed compared to longer, more complex sentences. The exception to this is with spelling errors where the predicted "yea" would give a high error rate when compared with the Spanish translation "ya", and "muy" would give an error when compared with the original elaborated "muyyy". Conjugational differences also led to increased word error rates. Perhaps a different metric would be more apt to analyze the quality of the constructed sentence.

4 Methodology

We aim to extend this approach to low-resource languages like Tulu.

4.1 Data Collection

In order to accomplish Tulu ASR, we will need access to Tulu audio with textual English and Kannada labels. Since there is no large public data set of Tulu audio and its corresponding translations, we will curate a data set by collecting Tulu movies with subtitles. There are several Tulu full-length movies and short films on YouTube with English subtitles available. We will use VideoSubFinder to split our video into frames, extract the text via OCR, and ultimately compile all the subtitles into a single line-separated text file. We can then use this for our model. We aim to collect 10 hours to Tulu data with English translations in this way. For any additional data we need, we will crawl YouTube for more videos using cutting edge techniques in conjunction with researcher Shinji Watanabe, and we will leverage data from Tulu translation dictionaries and translated Tulu songs described in section 2.4.

4.2 Approach

Our approach is based on the following modular pipeline.

$$\hat{\phi} = \arg\max_{\phi,f} P(x|\psi)P(\psi|\phi)P(\phi|f)P(f|e) \quad (3)$$

where ψ is phones corresponding to audio x, and all other variables are as defined in equation 2. We model $P(x|\psi)$ using Bayes' rule and a language-universal phone recognizer, Allosaurus [Li *et al.*, 2020]. (See section 2.2 for details.) This model has an intuitive API and Python-based application for phone recognition.

4.3 Mapping Phones to Phonemes

Though researchers have explored mapping phones to phonemes for ASR (see section 2.3), the problem of applying these mappings to new or unseen languages is an unsolved problem. We propose a novel method to map from predicted language universal phones to Tulu phonemes, which we will use to model $P(\psi|\phi)$. Allograph WFST models [Yan *et al.*, 2021], learn mappings from phones to language-specific phonemes. The Tulu phoneme inventory available on PHOIBLE[6] lists 37 Tulu phonemes and contains information about each. Using this resource, we will label a small portion of our Tulu audio data set (see section 4.1) with Tulu phonemes. Afterward we will pair these labels with phones predicted from the audio by the Allosaurus model [Li *et al.*, 2020]. With this small collection of aligned phones and Tulu phonemes, we will finetune Allographs. We have two possible approaches for this finetuning: (1) differential existing Allographs and update them for a phoneme prediction task, and (2) incorporate new language mappings to a larger Allograph via pruning. This can be accomplished by incorporating a fully-connected allophone graph for the target language into a well-trained multilingual Allograph, then pruning to acheive sparse phone-to-phoneme mappings for the new language.

4.4 Adaptation to New Languages

After applying this method for Tulu phoneme recognition, we will apply the best-peforming configuration to a small data set for the endangered language Pastaza Quichua. This set contains audio and video files with English transcription, glossing, and phonetic spellings, which will facilitate phoneme labeling if necessary.

5 Anticipated Difficulties

We anticipate a bottleneck in transcribing phonemes for Tulu audio data. We will likely need to perform a survey, pay transcribers, and develop general transformation rules using PHOIBLE for which a WFST can learn the likelihood.

[6] https://phoible.org/

6 Infrastructure

We intend to train models on Google Colab and Carnegie Mellon University's Linguistics Lab servers, as well as our own personal machinery.

References

Oliver Adams, Graham Neubig, Trevor Cohn, and Steven Bird. Learning a Translation Model from Word Lattices. In *Proc. Interspeech 2016*, pages 2518–2522, 2016.

Oliver Adams, Graham Neubig, Trevor Cohn, Steven Bird, Quoc Truong Do, and Satoshi Nakamura. Learning a lexicon and translation model from phoneme lattices. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2377–2382, Austin, Texas, November 2016. Association for Computational Linguistics.

P. J. Antony, Hemant B. Raj, B. S. Sahana, Dimple Sonal Alvares, and Aishwarya Raj. Morphological analyzer and generator for tulu language: a novel approach. In *ICACCI '12*, 2012.

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. An attentional model for speech translation without transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 949–959, San Diego, California, June 2016. Association for Computational Linguistics.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R. Mortensen, Graham Neubig, Alan W Black, and Florian Metze. Universal phone recognition with a multilingual allophone system, 2020.

Coskun Mermer and Murat Saraclar. Bayesian word alignment for statistical machine translation. In *ACL (Short Papers)*, pages 182–187, 2011.

Alexis Michaud, Eric Castelli, et al. Towards the automatic processing of yongning na (sino-tibetan): developing a'light'acoustic model of the target language and testing'heavyweight'models from five national languages. In *4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2014)*, pages 153–160, 2014.

Graham Neubig, Masato Mimura, Shinsuke Mori, and Tatsuya Kawahara. Bayesian learning of a language model from continuous speech. *IEICE Trans. Inf. Syst.*, 95-D:614–625, 2012.

Matthias Paulik and Alex Waibel. Training speech translation from audio recordings of interpreter-mediated communication. *Comput. Speech Lang.*, 27(2):455–474, February 2013.

Joris Pelemans, Tom Vanallemeersch, Kris Demuynck, Hugo Van hamme, and Patrick Wambacq. Efficient language model adaptation for automatic speech recognition of spoken translations. 09 2015.

B.S. Shivakumar. A wordnet for tulu. 2010.

E. Vidal, F. Casacuberta, L. Rodriguez, J. Civera, and C.D.M. Hinarejos. Computer-assisted translation using speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(3):941–951, 2006.

Brian Yan, Siddharth Dalmia, David R. Mortensen, Florian Metze, and Shinji Watanabe. Differentiable allophone graphs for language-universal speech recognition, 2021.

li Zezhong, Hideto Ikeda, and Junichi Fukumoto. Bayesian word alignment and phrase table training for statistical machine translation. *IEICE Transactions on Information and Systems*, E96.D:1536–1543, 07 2013.