Learning a Lexicon and Translation Model from Phoneme Lattices

Oliver Adams,[♠♡] **Graham Neubig,**[♣♡] **Trevor Cohn,**[♠]
Steven Bird,[♠] **Quoc Truong Do,**[♡] **Satoshi Nakamura**[♡]
[♠]The University of Melbourne, Australia
[♣]Carnegie Mellon University, Pittsburgh, PA, USA
[♡]Nara Institute of Science and Technology, Japan

Abstract

Language documentation begins by gathering speech. Manual or automatic transcription at the word level is typically not possible because of the absence of an orthography or prior lexicon, and though manual phonemic transcription is possible, it is prohibitively slow. On the other hand, translations of the minority language into a major language are more easily acquired. We propose a method to harness such translations to improve automatic phoneme recognition. The method assumes no prior lexicon or translation model, instead learning them from phoneme lattices and translations of the speech being transcribed. Experiments demonstrate phoneme error rate improvements against two baselines and the model's ability to learn useful bilingual lexical entries.

1 Introduction

Most of the world's languages are dying out and have little recorded data or linguistic documentation (Austin and Sallabank, 2011). It is important to adequately document languages while they are alive so that they may be investigated in the future. Language documentation traditionally involves one-on-one elicitation of speech from native speakers in order to produce lexicons and grammars that describe the language. However, this does not scale: linguists must first transcribe the speech phonemically as most of these languages have no standardized orthography. This is a critical bottleneck since it takes a trained linguist about 1 hour to transcribe the phonemes of 1 minute of speech (Do et al., 2014).

Smartphone apps for rapid collection of bilingual data have been increasingly investigated (De Vries et al., 2011; De Vries et al., 2014; Reiman, 2010; Bird et al., 2014; Blachon et al., 2016). It is common for these apps to collect speech segments paired with spoken translations in another language, making spoken translations quicker to obtain than phonemic transcriptions.

We present a method to improve automatic phoneme transcription by harnessing such bilingual data to learn a lexicon and translation model directly from source phoneme lattices and their written target translations, assuming that the target side is a major language that can be efficiently transcribed.[1] A Bayesian non-parametric model expressed with a weighted finite-state transducer (WFST) framework represents the joint distribution of source acoustic features, phonemes and latent source words given the target words. Sampling of alignments is used to learn source words and their target translations, which are then used to improve transcription of the source audio they were learnt from. Importantly, the model assumes no prior lexicon or translation model.

This method builds on work on phoneme translation modeling (Besacier et al., 2006; Stüker et al., 2009; Stahlberg et al., 2012; Stahlberg et al., 2014; Adams et al., 2015; Duong et al., 2016), speech translation (Casacuberta et al., 2004; Matusov et al., 2005), computer-aided translation, (Brown et al., 1994; Vidal et al., 2006; Khadivi and Ney, 2008; Reddy and Rose, 2010; Pelemans et al., 2015), translation modeling from automatically transcribed

[1]Code is available at https://github.com/oadams/latticetm.

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2377–2382,
Austin, Texas, November 1-5, 2016.

speech (Paulik and Waibel, 2013), word segmentation and translation modeling (Chang et al., 2008; Dyer, 2009; Nguyen et al., 2010; Chen and Xu, 2015), Bayesian word alignment (Mermer et al., 2013; Zezhong et al., 2013) and language model learning from lattices (Neubig et al., 2012). While we previously explored learning a translation model from word lattices (Adams et al., 2016), in this paper we extend the model to perform unsupervised word segmentation over phoneme lattices in order to improve phoneme recognition.

Experiments demonstrate that our method significantly reduces the phoneme error rate (PER) of transcriptions compared with a baseline recogniser and a similar model that harnesses only monolingual information, by up to 17% and 5% respectively. We also find that the model learns meaningful bilingual lexical items.

2 Model description

Our model extends the standard automatic speech recognition (ASR) problem by seeking the best phoneme transcription $\hat{\boldsymbol{\phi}}$ of an utterance in a joint probability distribution that incorporates acoustic features $\boldsymbol{x}$, phonemes $\boldsymbol{\phi}$, latent source words $\boldsymbol{f}$ and observed target transcriptions $\boldsymbol{e}$:

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi},\boldsymbol{f}}{\operatorname{argmax}} P(\boldsymbol{x}|\boldsymbol{\phi})P(\boldsymbol{\phi}|\boldsymbol{f})P(\boldsymbol{f}|\boldsymbol{e})\,, \qquad (1)$$

assuming a Markov chain of conditional independence relationships (bold symbols denote utterances as opposed to tokens). Deviating from standard ASR, we replace language model probabilities with those of a translation model, and search for phonemes instead of words. Also, no lexicon or translation model are given in training.

2.1 Expression of the distribution using finite-state transducers

We use a WFST framework to express the factors of (1) since it offers computational tractability and simple inference in a clear, modular framework. Figure 1 uses a toy German–English error resolution example to illustrate the components of the framework: a phoneme lattice representing phoneme uncertainty according to $P(\boldsymbol{x}|\boldsymbol{\phi})$; a lexicon that transduces phoneme substrings $\boldsymbol{\phi}_s$ of $\boldsymbol{\phi}$ to source tokens f according to $P(\boldsymbol{\phi}_s|f)$; and a lexical translation model representing $P(f|e)$ for each e in the written translation. The composition of these components is also shown at the bottom of Figure 1, illustrating how would-be transcription errors can be resolved. This framework is reminiscent of the WFST framework used by Neubig et al. (2012) for lexicon and language model learning from monolingual data.

2.2 Learning the lexicon and translation model

Because we do not have knowledge of the source language, we must learn the lexicon and translation model from the phoneme lattices and their written translation. We model lexical translation probabilities using a Dirichlet process. Let A be both the transcription of each source utterance $\mathbf{f}$ and its word alignments to the translation $\mathbf{e}$ that generated them. The conditional posterior can be expressed as:

$$P(f|e;A) = \frac{c_A(f,e) + \alpha P_0(f)}{c_A(e) + \alpha}\,, \qquad (2)$$

where $c_A(f,e)$ is a count of how many times f has aligned to e in A and $c_A(e)$ is a count of how many times e has been aligned to; P_0 is a base distribution that influences how phonemes are clustered; and α determines the emphasis on the base distribution.

In order to express the Dirichlet process using the WFST components, we take the union of the lexicon with a *spelling model* base distribution that consumes phonemes $\phi_i \ldots \phi_j$ and produces a special $\langle\text{unk}\rangle$ token with probability $P_0(\phi_i \ldots \phi_j)$. This $\langle\text{unk}\rangle$ token is consumed by a designated arc in the translation model WFST with probability $\frac{\alpha}{c_A(e)+\alpha}$, yielding a composed probability of $\frac{\alpha P_0(f)}{c_A(e)+\alpha}$. Other arcs in the translation model express the probability $\frac{c_A(f,e)}{c_A(e)+\alpha}$ of entries already in the lexicon. The sum of these two probabilities equates to (2).

As for the spelling model P_0, we consider three distributions and implement WFSTs to represent them: a geometric distribution, *Geometric(γ)*, a Poisson distribution, *Poisson(λ)*,[2] and a 'shifted' geometric distribution, *Shifted(α, γ)*. The shifted geometric distribution mitigates a shortcoming of the geometric distribution whereby words of length 1 have the highest probability. It does so by having

[2]While the geometric distribution can be expressed recursively, we cap the number of states in the Poisson WFST to 100.

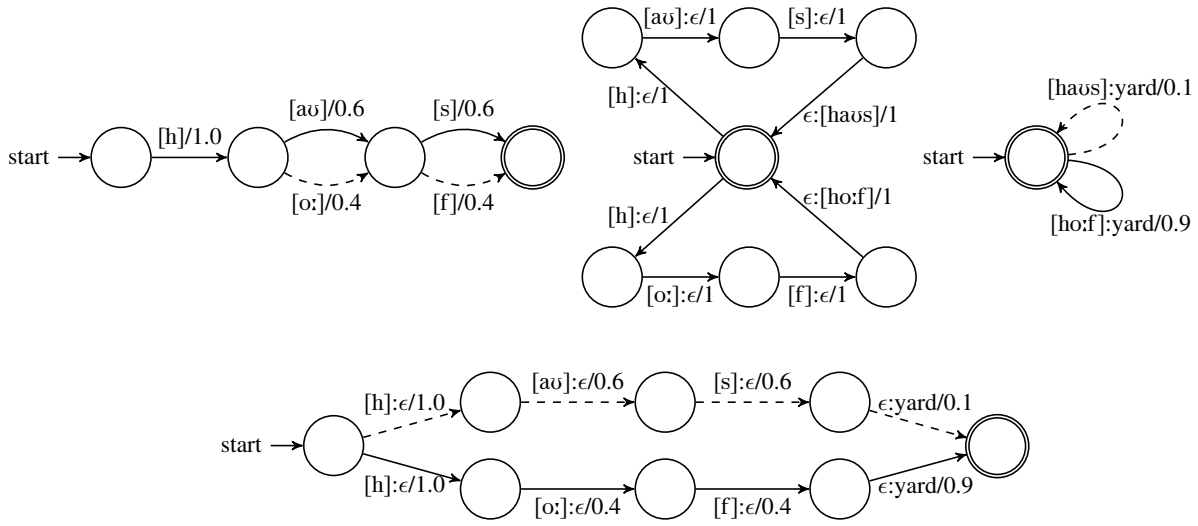

Figure 1: Top left to right: the phoneme lattice, the lexicon, and the translation model. Bottom: the resulting composed WFST. Given an English translation *'yard'*, the most likely transcription is corrected to [hoːf] (*'Hof'*) in the composed WFST, while in the original phoneme lattice it is [haʊs] (*'Haus'*). Solid edges represent most likely paths.

another parameter α that specifies the probability of a word of length 1, with the remaining probability mass distributed geometrically. All phonemes types are treated the same in these distributions, with uniform probability.

2.3 Inference

In order to determine the translation model parameters as described above, we require the alignments A. We sample these proportionally to their probability given the data and our prior, in effect integrating over all parameter configurations T:

$$P(A|\mathcal{X};\alpha,P_0) = \int_T P(A|\mathcal{X},T)P(T;\alpha,P_0)dT\,, \tag{3}$$

where $\mathcal{X}$ is our dataset of source phoneme lattices paired with target sentences.

This is achieved using blocked Gibbs sampling, with each utterance constituting one block. To sample from WFSTs, we use *forward-filtering/backward-sampling* (Scott, 2002; Neubig et al., 2012), creating forward probabilities using the forward algorithm for hidden Markov models before *backward-sampling* edges proportionally to the product of the forward probability and the edge weight.[3]

[3]No Metropolis-Hastings rejection step was used.

3 Experimental evaluation

We evaluate the lexicon and translation model by their ability to improve phoneme recognition, measuring phoneme error rate (PER).

3.1 Experimental setup

We used less than 10 hours of English–Japanese data from the BTEC corpus (Takezawa et al., 2002), comprised of spoken utterances paired with textual translations. This allows us to assess the approach assuming quality acoustic models. We used acoustic models similar to Heck et al. (2015) to obtain source phoneme lattices. Gold phoneme transcriptions were obtained by transforming the text with pronunciation lexicons and, in the Japanese case, first segmenting the text into tokens using KyTea (Neubig et al., 2011).

We run experiments in both directions: English–Japanese and Japanese–English (*en–ja* and *ja–en*), while comparing against three settings: the ASR 1-best path uninformed by the model (*ASR*); a monolingual version of our model that is identical except without conditioning on the target side (*Mono*); and the model applied using the source language sentence as the target (*Oracle*).

We tuned on the first 1,000 utterences (about 1 hour) of speech and trained on up to 9 hours of the

	English (en)			Japanese (ja)		
	Mono	–ja	Oracle	Mono	–en	Oracle
ASR		22.1			24.3	
Vague	17.7	18.5	17.2	21.5	20.8	21.6
Shifted	17.4	16.9	16.6	21.2	20.1	20.2
Poisson	17.3	17.2	16.8	21.3	20.1	20.8

Table 1: Phoneme error rates (percent) when training on 9 hours of speech, averaged over 4 runs.

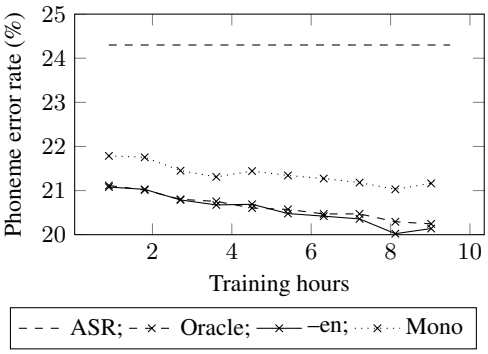

Figure 2: Japanese phoneme error rates using the *shifted* geometric prior when training data is scaled up from 1–9 hours, averaged over 3 runs.

remaining data.[4] Only the oracle setup was used for tuning, with Geometric(0.01) (taking the form of a *vague* prior), Shifted(10^{-5}, 0.25) and Poisson(7) performing best.

3.2 Results and Discussion

Table 1 shows en–ja and ja–en results for all methods with the full training data. Figure 2 shows improvements of ja–en over both the ASR baseline and the Mono method as the training data increases, with translation modeling gaining an increasing advantage with more training data.

Notably, English recognition gains less from using Japanese as the target side (en–ja) than the other way around, while the 'oracle' approach for Japanese recognition, which also uses Japanese as the target, underperforms ja–en. These observations suggest that using the Japanese target is less helpful, likely explained by the fine-grained morphological segmentation we used, making it harder for the model to relate source phonemes to target tokens.

The vague geometric prior significantly underperforms the other priors. In the en–ja/vague case, the model actually underperforms its monolingual counterpart. The vague prior biases slightly towards fine-grained English source segmentation, with words of length 1 most common. In this case, fine-grained Japanese is also used as the target which results in most lexical entries arising from uninformative alignments between single English phonemes and Japanese syllables, such as *[t]*⇔す. For similar reasons, the shifted geometric prior gains an advantage over Poisson, likely because of its ability to even further penalize single-phoneme lexical items, which regularly end up in all lexicons anyway due to their combinatorical advantage when sampling.

While many bilingual lexical entries are correct, such as *[wʌn]*⇔一 (*'one'*), most are not. Some have segmentation errors *[liːz]*⇔ください (*'please'*); some are correctly segmented but misaligned to commonly co-occurring words *[wʌt]*⇔時 (*'what'* aligned to *'time'*); others do not constitute individual words, but morphemes aligned to common Japanese syllables *[iːŋ]*⇔く (*'-ing'*); others still align multi-word units correctly *[haʊmʌtʃ]*⇔いくら (*'how much'*). Note though that entries such as those listed above capture information that may nevertheless help to reduce phoneme transcription errors.

4 Conclusion and Future Work

We have demonstrated that a translation model and lexicon can be learnt directly from phoneme lattices in order to improve phoneme transcription of those very lattices.

One of the appealing aspects of this modular framework is that there is much room for extension and improvement. For example, by using adaptor grammars to encourage syllable segmentation (Johnson, 2008), or incorporating language model probabilities in addition to our translation model probabilities (Neubig et al., 2012).

We assume a good acoustic model with phoneme error rates between 20 and 25%. In a language documentation scenario, acoustic models for the low-resource source language won't exist. Future work should use a universal phoneme recognizer or acoustic model of a similar language, thus making a step towards true generalizability.

[4] A 1 hour subset was used for PER evaluation.

Acknowledgments

We gratefully acknowledge support from the DARPA LORELEI program.

References

Oliver Adams, Graham Neubig, Trevor Cohn, and Steven Bird. 2015. Inducing bilingual lexicons from small quantities of sentence-aligned phonemic transcriptions. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2015)*, Da Nang, Vietnam.

Oliver Adams, Graham Neubig, Trevor Cohn, and Steven Bird. 2016. Learning a translation model from word lattices. In *17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016)*, San Francisco, California, USA.

Peter Austin and Julia Sallabank. 2011. *The Cambridge Handbook of Endangered Languages*. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

Laurent Besacier, Bowen Zhou, and Yuqing Gao. 2006. Towards speech translation of non written languages. In *2006 IEEE Spoken Language Technology Workshop (SLT 2006)*, pages 222–225, Palm Beach, Aruba.

Steven Bird, Florian R Hanke, Oliver Adams, and Haejoong Lee. 2014. Aikuma: A mobile app for collaborative language documentation. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–5, Baltimore, Maryland, USA.

David Blachon, Elodie Gauthier, Laurent Besacier, Guy-Noël Kouarata, Martine Adda-Decker, and Annie Rialland. 2016. Parallel speech collection for under-resourced language studies using the lig-aikuma mobile device app. *Procedia Computer Science*, 81:61–66.

Peter F Brown, Stanley F Chen, Stephen A Della Pietra, Vincent J Della Pietra, Andrew S Kehler, and Robert L Mercer. 1994. Automatic speech recognition in machine-aided translation. *Computer Speech & Language*, 8(3):177–187.

Francisco Casacuberta, Hermann Ney, Franz Josef Och, Enrique Vidal, Juan Miguel Vilar, Sergio Barrachina, Ismael García-Varea, David Llorens, César Martínez, Sirko Molau, and Others. 2004. Some approaches to statistical and finite-state speech-to-speech translation. *Computer Speech & Language*, 18(1):25–47.

Pi-Chuan Chang, Michel Galley, and Christopher D Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation (WMT 2008)*, pages 224–232, Columbus, Ohio, USA.

Wei Chen and Bo Xu. 2015. Semi-supervised Chinese word segmentation based on bilingual information. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pages 1207–1216, Lisbon, Portugal.

Nic J De Vries, Jaco Badenhorst, Marelie H Davel, Etienne Barnard, and Alta De Waal. 2011. Woefzela - an open-source platform for ASR data collection in the developing world. In *12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011)*, pages 3177–3180, Florence, Italy.

Nic J De Vries, Marelie H Davel, Jaco Badenhorst, Willem D Basson, Febe De Wet, Etienne Barnard, and Alta De Waal. 2014. A smartphone-based ASR data collection tool for under-resourced languages. *Speech Communication*, 56:119–131.

Thi-Ngoc-Diep Do, Alexis Michaud, and Eric Castelli. 2014. 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, St Petersburg, Russia.

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. 2016. 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 (NAACL HLT 2016)*, pages 949–959, San Diego, California, USA.

Chris Dyer. 2009. Using a maximum entropy model to build segmentation lattices for MT. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2009)*, pages 406–414, Boulder, Colorado, USA.

M Heck, Q T Do, S Sakti, G Neubig, T Toda, and S Nakamura. 2015. The NAIST ASR system for IWSLT 2015. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2015)*, Da Nang, Vietnam.

Mark Johnson. 2008. Unsupervised word segmentation for Sesotho using adaptor grammars. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology (SIGMORPHON 2008)*, pages 20–27, Columbus, Ohio, USA.

Shahram Khadivi and Hermann Ney. 2008. Integration of speech recognition and machine translation in computer-assisted translation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(8):1551–1564.

Evgeny Matusov, Stephan Kanthak, and Hermann Ney. 2005. On the integration of speech recognition and statistical machine translation. In *6th Interspeech 2005 and 9th European Conference on Speech Communication and Technology (INTERSPEECH 2005)*, pages 3177–3180, Lisbon, Portugal.

Coskun Mermer, Murat Saraçlar, and Ruhi Sarikaya. 2013. Improving statistical machine translation using Bayesian word alignment and Gibbs sampling. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(5):1090–1101.

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2 (ACL HLT 2011)*, pages 529–533, Portland, Oregon, USA.

Graham Neubig, Masato Mimura, and Tatsuya Kawahara. 2012. Bayesian learning of a language model from continuous speech. *IEICE TRANSACTIONS on Information and Systems*, 95(2):614–625.

ThuyLinh Nguyen, Stephan Vogel, and Noah A Smith. 2010. Nonparametric word segmentation for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 815–823, Beijing, China.

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

Joris Pelemans, Tom Vanallemeersch, Kris Demuynck, Patrick Wambacq, and Others. 2015. Efficient language model adaptation for automatic speech recognition of spoken translations. In *16th Annual Conference of the International Speech Communication Association (INTERSPEECH 2015)*, pages 2262–2266, Dresden, Germany.

Aarthi Reddy and Richard C Rose. 2010. Integration of statistical models for dictation of document translations in a machine-aided human translation task. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(8):2015–2027.

D Will Reiman. 2010. Basic oral language documentation. In *Language Documentation & Conservation*, pages 254–268.

Steven L Scott. 2002. Bayesian methods for hidden Markov models. *Journal of the American Statistical Association*, pages 337–351.

Felix Stahlberg, Tim Schlippe, Sue Vogel, and Tanja Schultz. 2012. Word segmentation through cross-lingual word-to-phoneme alignment. In *2012 IEEE Workshop on Spoken Language Technology (SLT 2012)*, pages 85–90, Miami, Florida, USA.

Felix Stahlberg, Tim Schlippe, Stephan Vogel, and Tanja Schultz. 2014. Word segmentation and pronunciation extraction from phoneme sequences through cross-lingual word-to-phoneme alignment. *Computer Speech & Language*, pages 234–261.

Sebastian Stüker, Laurent Besacier, and Alex Waibel. 2009. Human translations guided language discovery for ASR systems. In *10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009)*, pages 3023–3026, Brighton, United Kingdom.

Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 147–152, Las Palmas, Canary Islands.

Enrique Vidal, Francisco Casacuberta, Luis Rodriguez, Jorge Civera, and Carlos D Martínez Hinarejos. 2006. Computer-assisted translation using speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(3):941–951.

L I Zezhong, Hideto Ikeda, and Junichi Fukumoto. 2013. Bayesian word alignment and phrase table training for statistical machine translation. *IEICE TRANSACTIONS on Information and Systems*, 96(7):1536–1543.