INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Learning a Translation Model from Word Lattices

Oliver Adams,[1] *Graham Neubig,*[2] *Trevor Cohn,*[1] *Steven Bird*[1]

[1]The University of Melbourne, Australia
[2]Nara Institute of Science and Technology, Japan
oadams@student.unimelb.edu.au, neubig@is.naist.jp,
tcohn@unimelb.edu.au, sbird@unimelb.edu.au

Abstract

Translation models have been used to improve automatic speech recognition when speech input is paired with a written translation, primarily for the task of computer-aided translation. Existing approaches require large amounts of parallel text for training the translation models, but for many language pairs this data is not available. We propose a model for learning lexical translation parameters directly from the word lattices for which a transcription is sought. The model is expressed through composition of each lattice with a weighted finite-state transducer representing the translation model, where inference is performed by sampling paths through the composed finite-state transducer. We show consistent word error rate reductions in two datasets, using between just 20 minutes and 4 hours of speech input, additionally outperforming a translation model trained on the 1-best path.
Index Terms: speech recognition, machine translation

1. Introduction

The quality of automatic speech recognition (ASR) can often be improved by harnessing textual translations [1, 2, 3]. One use case is computer-aided translation (CAT) with interfaces where a human translator speaks their translation of a written document. Although monolingual ASR alone may suffice for ergonomics or efficiency, performance gains have been demonstrated by using translation models (TMs) to help disambiguate the translator's speech. Thus, when distinguishing acoustically ambiguous candidates such as *'recognise speech'* from *'wreck a nice beach'*, translation model scores make it easier to disambiguate these utterances when paired with a written translation in another language.

Existing work requires externally sourced parallel text data, a scarce resource for most language pairs even when each language has substantial monolingual data. We propose a method for improving speech recognition accuracy by harnessing a written translation in another language, even when no parallel text corpora are available. This is achieved by training a translation model directly on word lattice data from utterances we seek to recognize paired with written translations in another language.

We use a generative model that assumes the acoustic signal and written translation are produced by some underlying word sequence we seek to recover. This model is expressed by composing a word lattice that expresses information from the ASR acoustic and language models with a weighted finite-state transducer (WFST) that expresses lexical translation probabilities constrained by the observed translation. These parameters are learnt by sampling paths (word sequences) through the composed WFST, together with their alignments to translations. A likely source sentence is recovered by finding the shortest path in the WFST.

In experiments on the Fisher and CALLHOME Spanish–English Speech Translation Corpus [4], we compare word error rates with those of ASR 1-best paths and a stronger baseline that trains an existing translation model on 1-best recognition results. We demonstrate reduced word error rates of 4.1% to 5.6% relative over the 1-best paths, and also show 2.3% to 2.4% relative improvement over the alternative model that uses parameters learnt using 1-best paths. These results indicate that the mere existence of translations of what is to be transcribed can help with ASR. Moreover, it shows promise for models of this type for computer-aided translation and also for speech recognition for low-resource languages, where neither translation nor recogniser technologies are currently adequate.

2. Related work

There has been extensive work on combining ASR and statistical machine translation (SMT) systems, with the work largely focused on coupling the systems for the problem of speech translation [5, 6, 7, 8]. There has also been a variety of work on using translation models to improve ASR performance [9, 10], which includes the popular CAT use case [11, 1]. Work has typically involved modifying language model probabilities in the ASR system [12, 13], and improving decision-making between the N-best hypotheses of ASR systems [14, 15]. Additionally, word lattice based approaches have also been pursued [2, 3]. The transcription of multiple streams of interpreted speech has also been addressed with the aid of machine translation [16, 17]. However, in all of these works the translation models are trained on substantial external written corpora such as European parliament proceedings or the Canadian Hansards.

There has been scarce work in the area of training translation models from speech data. Notably, [18, 19, 20] investigate using interpreted speech from European Parliament Plenary Sessions, first performing alignment of speech utterances before subsequent word-level alignment of their 1-best ASR hypotheses with a traditional word aligner.

Our approach differs in that (a) it depends on no prior parallel text training data, and that (b) the translation model is trained directly from word lattices to harness more information than is available in the 1-best ASR hypothesis alone.

This approach uses techniques similar to those found in the Bayesian word alignment literature [21, 22, 23], though rather than sampling alignments between observed source and target

http://dx.doi.org/10.21437/Interspeech.2016-862

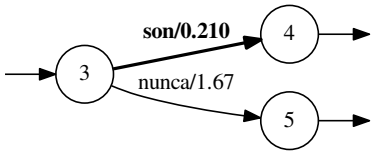

Figure 1: Part of a lattice for *'Ahá, pero una nunca sabe'* with translation *'Aha, but one never knows'*. Note that *'son'* is incorrectly given higher probability than *'nunca'*. Probabilities are expressed as negative log probabilities.

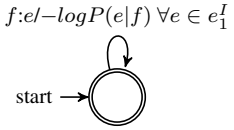

Figure 2: Reduced translation model template and hypothetical translation model parameters. Edges are added to the WFST only if the e is present in written translation.

word sequences, we sample paths through the source word lattice jointly with alignments to the target (translation) word sequences.

This approach is also similar to [24], where a lexicon and language model are learnt directly from phoneme lattices. However, rather than composing a phoneme lattice with a lexicon and a language model WFST, we compose a word lattice with a WFST representing a translation model.

3. Model description

3.1. ASR lattices

ASR is characterized by the search problem

$$\hat{f_1^J} = \underset{f_1^J}{\operatorname{argmax}}\, P(x_1^T|f_1^J)P(f_1^J) \tag{1}$$

where f_1^J represents an unobserved sequence of words $f_1 \dots f_J$ that produced the sequence of observed acoustic features $x_1 \dots x_T$, and $\hat{f_1^J}$ is our best guess of those words.

An ASR lattice encodes multiple ASR hypotheses, as shown in Figure 1 where each edge corresponds to a word f_i. The acoustic model (AM) and language model (LM) probabilities $P(x_1^T|f_1^J)$ and $P(f_1^J)$ are captured by the weights of the edges.

For any path f_1^J through the lattice, its probability can simply be determined with $P(f_1^J) = \prod_{j=1}^{J} P_L(f_j)$, where $P_L(f_j)$ is the probability of the ith edge in that path of the lattice. The most likely path can be determined by finding the shortest path through the lattice in question using Dijkstra's algorithm, if probabilities are represented as negative log probabilities.

3.2. Proposed model

The proposed model also uses translation models to aid in ASR by incorporating additional information in the form of an observed sequence e_1^I of translated words

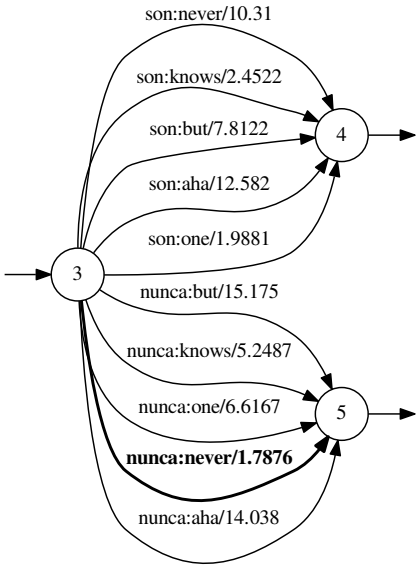

Figure 3: Part of the lattice composed with the translation model WFST. Each edge for a given Spanish word is replaced with a set of edges that transduce to different English words with probabilities re-weighted by the translation model. Note that *'nunca'* is now correctly given more weight than *'son'*, unlike in Figure 1.

$$\begin{aligned}\hat{f_1^J} &= \underset{f_1^J}{\operatorname{argmax}}\, P(f_1^J|x_1^T, e_1^I) \\ &= \underset{f_1^J}{\operatorname{argmax}}\, P(e_1^I, x_1^T, f_1^J).\end{aligned} \tag{2}$$

Assuming conditional independence of x_1^T and e_1^I given f_1^J, the problem can be factorized to

$$\hat{f_1^J} = \underset{f_1^J}{\operatorname{argmax}}\, P(e_1^I|f_1^J)P(x_1^T|f_1^J)P(f_1^J) \tag{3}$$

This problem can be reduced to a similar shortest-path problem with inference of f_1^J limited to the paths that occur in the original lattice. This is done by composing our original lattices with a WFST that represents translation model probabilities, as shown in Figure 2. The resulting composition of the final lattice and the constrained translation model can be seen in Figure 3.[1]

In this framework, each path represents a sequence of source tokens f_1^J and the sequence of target words they are aligned to, $e_{a_1} \dots e_{a_J}$,[2] with probability:

$$\begin{aligned}P(e_1^I, x_1^T, f_1^J) &\approx P(e_{a_1} \dots e_{a_J}|f_1^J)P(x_1^T|f_1^J)P(f_1^J) \\ &= \prod_{j=1}^{J} \left\{P(e_{a_j}|f_j)P(f_j)\right\} P(x_1^T|f_1^J)\end{aligned} \tag{4}$$

[1] For the examples in the figures and this formulation, we disregard the possibility of the null token, which we discuss in Section 5.

[2] In our implementation, we employ an optimisation by considering alignment only to unique word *types* on the target side, rather than tokens. This is possible as the probability will be the same for identical tokens.

In this case the shortest path corresponds to determining the most likely source f_1^J and the alignments a_1^J. Since distinct alignment paths may have the same source form f_1^J, $\hat{f}_1^J$ is most accurately found by marginalizing over the alignments: $\hat{f}_1^J = \text{argmax}_{f_1^J} \sum_{a_1^J} P(f_1^J, e_{a_1} \dots e_{a_J})$. However, for reasons of computational tractability we simply approximate $\hat{f}_1^J$ with the source side of the most likely path.

4. Parameter learning

We now turn to the task of determining parameters which will allow us to the find the form of f_1^J as discussed above. We model these parameters using a Dirichlet distribution:

$$P(e|f;\alpha) = \frac{c_{e,f} + \alpha P_{base}(e|f)}{\sum_{e'} c_{e',f} + \alpha} \quad (5)$$

where e' are target tokens in e_1^I, $c_{e,f}$ is a count of how many times f has aligned to e in the rest of the dataset (holding out this current instance), P_{base} is a uniform prior, and α is the prior's strength. This requires that we have a set of source tokens and alignments $\mathcal{A} = \{\langle f_1^{J_n}, a_1^{J_n}\rangle | n \in N\}$ where J_n is the number of edges in a sampled path from the nth WFST composition.

We want to base these parameters on alignments $\mathcal{A}$ with a probability proportional to $\mathcal{A}$'s ability to explain the data:

$$P(\mathcal{A}|\mathcal{X};\alpha) = \int_T P(\mathcal{A}|\mathcal{X},T)P(T;\alpha)dT \quad (6)$$

where $\mathcal{X}$ represents all the observed data, including word lattices and written translations, and T represents a translation model that we assume prior information about and whose parameters we integrate over.

Draws from the above distribution over $\mathcal{A}$ are approximated using blocked Gibbs sampling, where each block corresponds to one of the composed WFSTs. The sampling of paths through these composed WFSTs can be achieved using the method of *forward-filtering/backward-sampling* as described in [24]. This sampling method first computes forward probabilities in the same way the forward-backward algorithm for hidden Markov models does. It then samples paths backwards from the end of the WFST using these forward probabilities to yield a path with probability proportional to the total probability of the edges in the path.

After sampling a path consisting of lexical alignments, the counts of those lexical alignments are added to the cache used to calculate the Dirichlet posterior of e given f as per (5) before the next WFST is constructed and sampled from. With repeated sampling the alignment set $\mathcal{A}$ is drawn approximately from the distribution in (6). Sampling sets of alignments n times, we use these alignment sets $\mathcal{A}_1 \dots \mathcal{A}_n$ to create a set of point estimates $T_1 \dots T_n$ for T. We then average these parameters to create a final expected T for the purposes of decoding using the approach of Section 3.

5. Variations on parameter formulation

In the previous formulation of (4) we discussed using the conditional direction of $P(e|f)$ as the translation model parameters. A problem with using $P(e|f)$ is that values of e with a higher marginal probability $P(e)$ tend to have a higher conditional probability $P(e|f)$. This problem makes itself especially clear when permitting null tokens on the English side, as it leads to degenerate alignments where most f_i end up aligning to the null token since it is present in every sentence.

	1-best TM	Lattice TM	
		$\alpha, \lambda = 1$	$\alpha, \lambda =$ best
$P(e\|f)$	0.555	0.559	0.556
$P(f\|e)$	0.552	0.542	0.541
$\frac{P(e\|f)}{sum_{e'} P(e'\|f)}$	0.568	0.574	0.570
$\frac{P(f\|e)}{sum_{f'} P(f'\|e)}$	0.547	**0.539**	**0.539**

Table 1: Parameter variations for tuning on the CALLHOME training set. ASR 1-best accuracy is 0.569.

Alternative formulations include use of $P(f|e)$ instead for both training and testing, and/or the use of normalizations of these probabilities. Notably, we propose and test an approach that uses $\frac{P(f|e)}{\sum_{f'} P(f'|e)}$ where each f' is a token occurring in the original lattice.[3] The normalization denominator does not explicitly affect distinguishing between different source words in the WFST when sampling or decoding. However, it aids in aligning to the correct target word e by biasing towards alignments where f is most likely relative to its peers given e. Improving the alignments this way thus affects the translation model and, subsequently, the future paths chosen when sampling or decoding.

We also introduce a lattice weight λ during both training and testing. The contribution of original lattice probabilities from the acoustic model and language model against the translation model probabilities can be increased by simply multiplying the negative log probabilities by λ.

6. Experimental evaluation

6.1. Experimental setup

For the experiments we used the Fisher and CALLHOME Spanish–English Speech Translation Corpus [4], which conveniently offers Spanish word lattices and crowdsourced English translations. For training and testing we use the predefined test subsets of these corpora which are 213 minutes (39,978 words) for the Fisher corpus and 106 (18,792 words) for the CALLHOME corpus. The LDC human transcriptions [25, 26] are used as a gold standard against which to evaluate the ASR. Our preprocessing involved lowercasing all text, and removing punctuation from both the Spanish and English sides. We additionally removed from the corpus a small number of empty sentences and empty lattices.

To evaluate how harnessing the English translations can improve use of the Spanish word lattices, we evaluate the word error rate of the chosen path through the composed WFST against the LDC transcriptions. We compare our approach, which we refer to as *Lattice TM*, with a similar method where the translation model is instead trained from 1-best paths from the lattice using GIZA++ [27], which we refer to as *1-best TM*.

6.2. Tuning and choice of parameterization

We tuned two hyperparameters on the CALLHOME training set of approximately 14.5 hours: the lattice weight, λ, and the concentration parameter of the Dirichlet distributions, α. Tuning involved a simple grid search of λ, α over the values 0.5, 1, 2, and 4. We found no significant improvements beyond the de-

[3] This denominator does not equal 1 as all f' in the lattice constitute only a subset of the source vocabulary.

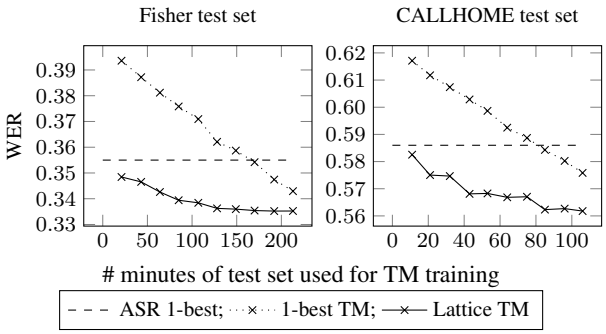

Figure 4: Word error rates on the test sets when training the translation model on different subsets of that test set.

	Fisher	CALLHOME
ASR 1-best	0.355	0.586
1-best TM	0.343	0.576
Lattice TM	**0.335**	**0.562**

Table 2: Word error rates when training and testing on the Fisher and CALLHOME test sets, of 213 minutes (39,978 words) and 106 minutes (18,792 words) respectively.

fault values of 1 for both parameters. Though values of λ at 4 began to degrade, the improvements were still significantly better than the ASR 1-best WER with these improvements robust for all hyperparameter combinations evaluated. With no strong motivation to deviate from the defaults, we left these parameters at $\lambda = 1$ and $\alpha = 1$ and evaluated on the test sets.

During tuning, we also evaluated the parameterizations discussed in 5. The best parameterization, $\frac{P(f|e)}{\sum_{f'} P(f'|e)}$, was used for the subsequent evaluation. Permitting null alignments in the translation model WFST reduced performance for all parameter variations, most notably $P(e|f)$. The model results presented permit no null alignments.

6.3. Results and discussion

Table 2 shows the results of unsupervised learning and evaluation across both of the test sets, with the 1-best TM outperforming the ASR baseline, but underperforming Lattice TM on both test sets. Figure 4 illustrates the change in performance when the models are restricted by being trained on subsets of the test sets. Training data is scaled up towards the natural unsupervised case where the training data comprises all the lattice data to be decoded. When training data is limited, the translation model trained on the 1-best path adversely affects performance, increasing the WER. In contrast, Lattice TM remains robust.

An example (used in Figure 1) from the CALLHOME test set epitomizes why the TM learnt from the lattice outperforms that learnt from the 1-best path. The English translation is *'aha, but one never knows'* and the gold transcription is *'Ahá, pero una nunca sabe'*.[4] The better path is the bottom one, choosing *'nunca'* over *'son'*. However, the 1-best path chooses *'son'*. Training a translation model from the erroneous 1-best path causes negative reinforcement, where the TM is even more likely to assign a high probability to *'son'* given *'never'*, and in the absence of sufficient training data, this has significant effect.

Since *'never'* and *'nunca'* are relatively frequently occurring in the test data, the 1-best TM actually assigns a reasonable lexical translation score to this pair, however this is not enough to overcome the lattice's bias and the reasonable probability of *'son'* given *'never'* learnt from the erroneous transcription.

[4]The gold transcription is notably unobtainable from the lattice. Reduced pruning of lattices are likely to further improve scores.

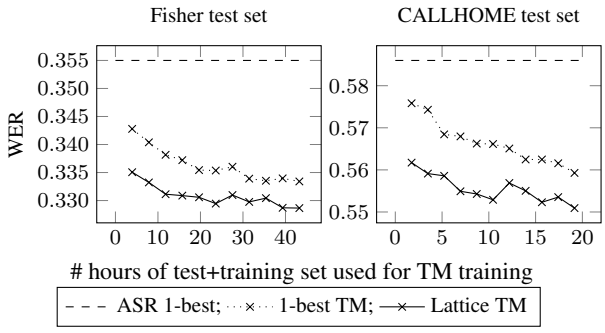

Figure 5: Word error rates when training data is scaled up to many hours.

It is also worth considering how the performance of these methods scale up to more data. Figure 5 presents evaluations on the same test sets, but permitting more training data from the respective corpora. It is interesting to note that Lattice TM continues to outperform the 1-best TM approach. This suggests that Lattice TM gains an advantage from the extra information encoded in the lattice beyond avoiding the negative reinforcement of the 1-best TM approach.

This method is very fast, with composition, sampling and caching for 1,000 utterances taking between 3 and 4 seconds on a single 1.80GHz Intel i7-4500U core. Running on the 213 minute Fisher evaluation set (Table 2) took less than 5 minutes, and scales roughly linearly with more training data.

7. Conclusion and future work

We have demonstrated that having a written translation in another language can help improve speech recognition even when no pre-trained translation model is available. This is achieved by training a translation model directly on the ASR word lattices paired with the written translation, in order to make the most of all information available in the lattice.

One natural setting for such an approach is for computer-aided translation of a small language for which there exists written data but no parallel corpora with the larger target language. However, since most languages have inadequate or no ASR technology and stand to gain the most from improved speech recognition systems, future work should also strive to reverse the role of the languages in this setup, addressing the speech of a small language paired with a written translation in a larger language. Such bilingual data can be collected using a tool such as *Aikuma* [28]. For this to work, an ASR system with a lexicon and language model needs to be trained (using a tool such as *Woefzela* [29]), or sidestep this need by working directly on the speech signal or phoneme lattices. There also exist many languages which have limited but detailed parallel data and comprehensive linguistic description. Such approaches may prove useful there by bootstrapping a translation model with the available glosses and improving ASR of the language with the methods described in this paper.

Finally, it is likely this work could all be extended to cope with interpreted speech using lattices on both the source and target sides, though increased computational complexity must be addressed along with the nuances associated with interpreted speech [20].

8. References

[1] P. F. Brown, S. F. Chen, S. A. Della Pietra, V. J. Della Pietra, A. S. Kehler, and R. L. Mercer, "Automatic speech recognition in machine-aided translation," *Computer Speech & Language*, vol. 8, no. 3, pp. 177–187, 1994.

[2] S. Khadivi and H. Ney, "Integration of speech recognition and machine translation in computer-assisted translation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 8, pp. 1551–1564, 2008.

[3] A. Reddy and R. C. Rose, "Integration of statistical models for dictation of document translations in a machine-aided human translation task," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 8, pp. 2015–2027, 2010.

[4] M. Post, G. Kumar, A. Lopez, D. Karakos, C. Callison-Burch, and S. Khudanpur, "Improved Speech-to-Text Translation with the Fisher and Callhome Spanish–English Speech Translation Corpus," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, dec 2013.

[5] E. Vidal, "Finite-state speech-to-speech translation," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 1. IEEE, 1997, pp. 111–114.

[6] E. Matusov, S. Kanthak, and H. Ney, "On the integration of speech recognition and statistical machine translation." in *INTERSPEECH*, 2005, pp. 3177–3180.

[7] H. Ney, "Speech translation: Coupling of recognition and translation," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 1. IEEE, 1999, pp. 517–520.

[8] F. Casacuberta, H. Ney, F. J. Och, E. Vidal, J. M. Vilar, S. Barrachina, I. García-Varea, D. Llorens, C. Martínez, S. Molau, and Others, "Some approaches to statistical and finite-state speech-to-speech translation," *Computer Speech & Language*, vol. 18, no. 1, pp. 25–47, 2004.

[9] E. Vidal, F. Casacuberta, L. Rodriguez, J. Civera, and C. D. M. Hinarejos, "Computer-assisted translation using speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 3, pp. 941–951, 2006.

[10] V. Alabau, L. Rodríguez-Ruiz, A. Sanchis, P. Martínez-Gómez, and F. Casacuberta, "On multimodal interactive machine translation using speech recognition," in *Proceedings of the 13th international conference on multimodal interfaces*. ACM, 2011, pp. 129–136.

[11] P. Isabelle, M. Dymetman, G. Foster, J.-M. Jutras, E. Macklovitch, F. Perrault, X. Ren, and M. Simard, "Translation analysis and translation automation," in *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing-Volume 2*. IBM Press, 1993, pp. 1133–1147.

[12] L. Rodriguez, A. Reddy, and R. Rose, "Efficient integration of translation and speech models in dictation based machine aided human translation," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4949–4952.

[13] J. Pelemans, T. Vanallemeersch, K. Demuynck, P. Wambacq, and Others, "Efficient language model adaptation for automatic speech recognition of spoken translations," in *Proceedings Interspeech 2015*, 2015, pp. 2262–2266.

[14] R. W. M. Ng, T. Hain, and T. Cohn, "Adaptation of lecture speech recognition system with machine translation output," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8401–8405.

[15] M. Paulik, S. Stüker, C. Fugen, T. Schultz, T. Schaaf, and A. Waibel, "Speech translation enhanced automatic speech recognition," in *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*. IEEE, 2005, pp. 121–126.

[16] J. Miranda, J. P. Neto, and A. W. Black, "Parallel combination of speech streams for improved ASR," *Proceedings of the Interspeech, Portland, USA*, 2012.

[17] ——, "Recovery of acronyms, out-of-lattice words and pronunciations from parallel multilingual speech," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 348–353.

[18] M. Paulik and A. Waibel, "Automatic translation from parallel speech: Simultaneous interpretation as MT training data," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 496–501.

[19] ——, "Spoken language translation from parallel speech audio: Simultaneous interpretation as SLT training data," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 5210–5213.

[20] ——, "Training speech translation from audio recordings of interpreter-mediated communication," *Computer Speech & Language*, vol. 27, no. 2, pp. 455–474, 2013.

[21] C. Mermer and M. Saraçlar, "Bayesian word alignment for statistical machine translation," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, 2011, pp. 182–187.

[22] C. Mermer, M. Saraçlar, and R. Sarikaya, "Improving statistical machine translation using Bayesian word alignment and Gibbs sampling," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 5, pp. 1090–1101, 2013.

[23] L. I. Zezhong, H. Ikeda, and J. Fukumoto, "Bayesian word alignment and phrase table training for statistical machine translation," *IEICE TRANSACTIONS on Information and Systems*, vol. 96, no. 7, pp. 1536–1543, 2013.

[24] G. Neubig, M. Mimura, and T. Kawahara, "Bayesian learning of a language model from continuous speech," *IEICE TRANSACTIONS on Information and Systems*, vol. 95, no. 2, pp. 614–625, 2012.

[25] B. Wheatley, "CALLHOME Spanish Transcripts LDC96T17," Web Download, Philadelphia, USA, 1996.

[26] D. Graff, S. Huang, I. Cartagena, K. Walker, and C. Cieri, "Fisher Spanish Transcripts LDC2010T04," Web Download, Philadelphia, USA, 2010.

[27] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[28] S. Bird, F. R. Hanke, O. Adams, and H. Lee, "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*. Baltimore, Maryland, USA: ACL, jun 2014, pp. 1–5.

[29] N. J. De Vries, J. Badenhorst, M. H. Davel, E. Barnard, and A. De Waal, "Woefzela - an open-source platform for ASR data collection in the developing world," in *INTERSPEECH*, 2011.