Inducing Bilingual Lexicons from Small Quantities of Sentence-Aligned Phonemic Transcriptions

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

[1]Department of Computing and Information Systems, The University of Melbourne, Australia
[2]Graduate School of Information Science, Nara Institute of Science and Technology, Japan

oadams@student.unimelb.edu.au, neubig@is.naist.jp, {t.cohn,sbird}@unimelb.edu.au

Abstract

We investigate induction of a bilingual lexicon from a corpus of phonemic transcriptions that have been sentence-aligned with English translations. We evaluate existing models that have been used for this purpose and report on two additional models, which demonstrate performance improvements. The first performs monolingual segmentation followed by alignment, while the second performs both tasks jointly. We show that monolingual and bilingual lexical entries can be learnt with high precision from corpora having just 1k–10k sentences. We explain how our results support the application of alignment algorithms to the task of documenting endangered languages.

1. Introduction

Language documentation involves the construction of text collections, lexicons and grammars in the interest of creating a record of a language for future linguistic, cultural and anthropological analysis. Traditional approaches to language documentation are labour-intensive, requiring much one-on-one time between a field linguist and the mother tongue speakers. Unfortunately, there aren't enough linguists to document the world's languages using these approaches before many of the approximately 7,000 languages die out.

There is a movement to increase the rate of data collection of endangered languages using cheap and widespread electronics to record speech in a more ad hoc manner [1, 2, 3, 4, 5], in an attempt to provide the field linguist with leverage to acquire data faster. This data is primarily audio, since most languages have no established written form and capturing audio is comparatively fast. Additionally, much of the data is bilingual, as an important aspect of the language documentation process is the construction of bilingual corpora and lexicons.

In this paper we consider the task of automatically learning monolingual and bilingual lexical items from unsegmented phonemic transcriptions of interleaved audio (segments of speech in one language along with spoken translations in another). Such transcriptions could arise from two scenarios. The first is when future philologists phonetically transcribe speech of a language post-mortem, without native speakers to assist in word segmentation. In such instances lexicon induction would aid in linguistic analysis of the language. The second is by instead employing automatic speech recognition technologies for the same task. In both cases lexicon induction could aid in bootstrapping automatic speech recognition (ASR) systems targeting the language's untranscribed audio. Note that we assume a transcription of the English translation, since English speech can be reliably and cheaply transcribed.

Previous work on bilingual lexicon induction using sentence-aligned corpora has focused primarily on large corpora of written text [6, 7, 8, 9]. However, bilingual lexicon induction applied to phonemically transcribed audio introduces problems, including the lack of word segmentation and the small quantities of data. There has been limited work on learning lexicons from phonemic transcriptions. [10, 11] take a first look at phoneme–word translation modeling, using traditional IBM Models [12] in order to determine alignments and applying heuristics to extract dictionaries. [13] propose Model 3P, which builds upon the generative story of IBM Model 3 by adding additional word length parameters and allowing it to significantly outperform the IBM models [14, 15, 16].

Building on this work, we investigate two models that haven't been considered in this context and demonstrate that they can outperform the models that have been considered. The first performs unsupervised word segmentation followed by word alignment. The second jointly performs word segmentation and alignment. Importantly, we evaluate the models on a data set that is significantly smaller than has been evaluated on previously, containing between just 1k and 10k sentences, corresponding to 13k and 132k words. This likely corresponds to something in the order of 1 to 10 hours of speech [17, 18, 5]. These quantities of data are realistic in the context of documentation of endangered languages, though the applicability of these techniques also applies more generally to low-resource languages that have no body of written resources.

We run experiments to assess the induced lexicons' precisions at k entries. We do this by applying the alignment models to a German–English corpus, using heuristics to extract lexical entries before having them manually annotated[1].

German was used since it permitted easier manual annotation of lexical entries than an endangered language. Although German and English are more closely related languages than language pairs encountered in linguistic fieldwork, modeling of the language pair is still complex due to varying word order between the languages and the morphological richness of German relative to English.

Results demonstrate that hundreds of bilingual lexical entries can be learnt with good precision, with the additional proposed methods outperforming Model 3P on a data set of 10k sentences. This offers promise of the technique's applicability in a language documentation context. Moreover, the majority of incorrect entries correspond to well-segmented, but misaligned, source words.

2. Translation Models

Our lexicon induction approach uses various phrase alignment techniques to segment sequences of phonemes into words and learn phrase tables. There are several methods for word segmentation in machine translation [19, 20, 21, 22, 16], but there has been limited application in a low resource context. In this paper we examine four representative methods to apply to parallel sentences comprised of source phoneme tokens and target words.

The first two, GIZA++ and Model 3P, have been investigated previously for the task of phoneme–word alignment [10, 14]. They are evaluated as a point of comparison for the latter two methods we demonstrate are effective for this task, which use unsupervised word segmentation (UWS) with GIZA++ and a Bayesian inversion transduction grammar (ITG) framework.

2.1. GIZA++

GIZA++ is the baseline that follows the standard statistical machine translation (SMT) pipeline of performing alignment with the IBM Models [12], as implemented in GIZA++ [23]. This approach to alignment was used in seminal work on phoneme–word alignment [10, 11]. The problem with this approach is that it attempts to capture relationships between individual foreign phonemes and English words, which is extremely difficult.

2.2. Model 3P

PISA[2] is an implementation of the Model 3P model of [13]. It builds upon the generative model of IBM Model 3 [12] by adding additional word length parameters (see Figure 1), allowing it to outperform traditional IBM models on phoneme–word alignment tasks. After initializing model parameters with learnt GIZA++ parameters, the PISA implementation of Model 3P uses a genetic algorithm to learn the parameters of the model.

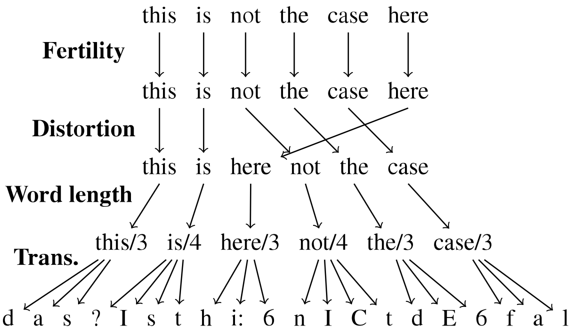
this is not the case here
Fertility
this is not the case here
Distortion
this is here not the case
Word length
Trans.
this/3 is/4 here/3 not/4 the/3 case/3
d a s ? I s t h i: 6 n I C t d E 6 f a l

Figure 1: The generative model of Model 3P.

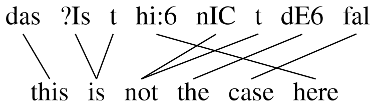

Figure 2: Monolingual segmentation of phonemes followed by alignment, as done in the UWS GIZA++ approach.

The additional word length parameters, distinct from the fertility parameters, allow Model 3P to learn latent word representations that would not be able to be captured in a direct phoneme–word mapping. This allows for better segmentation performance.

2.3. UWS GIZA++

UWS GIZA++ first performs unsupervised word segmentation using the Bayesian Pitman-Yor language model [24], as implemented in the tool pgibbs[3] [25]. Alignment is then performed between these phoneme sequences and the English words using GIZA++ (see Figure 2). This was hypothesized to be more appropriate than GIZA++ alone since it would result in breaking the foreign phoneme sequences into coarser tokens that translate better to English. Note that there is not an expectation that the word segmentation perform well with respect to what is considered a "word" in the given language. Instead, the key idea is that the segmenter breaks phonemes into frequently repeating units that capture more meaning than just using individual phonemes. Consider Figure 2: the erroneous segmentation nevertheless allows for accurate alignment after monolingual segmentation.

2.4. Bayes ITG

Bayes ITG performs joint word segmentation and alignment using the substring alignment model of [26], as implemented in pialign[4] [27]. Alignments are obtained through Bayesian learning of inversion transduction grammar trees [28], which

[1] These annotations will be released along with code for the lexicon induction.

[2] https://code.google.com/p/pisa

[3] http://github.com/neubig/pgibbs

[4] http://github.com/neubig/pialign

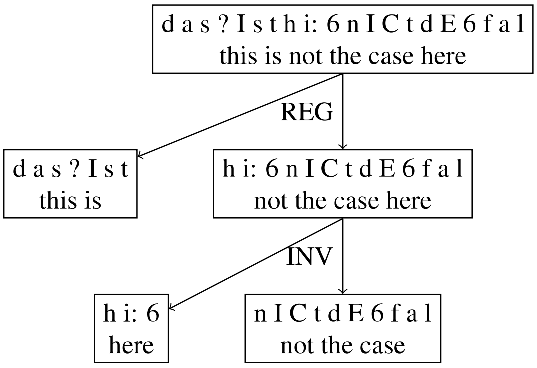
d a s ? I s t h i: 6 n I C t d E 6 f a l
this is not the case here
REG
d a s ? I s t
this is
h i: 6 n I C t d E 6 f a l
not the case here
INV
h i: 6
here
n I C t d E 6 f a l
not the case

Figure 3: An ITG tree structure learnt by pialign. Note that pialign forces alignments down to individual tokens, but the leaf nodes presented here represent alignments that were generated as single phrase by the model.

completely describe the sentence and its translation as a tree of aligned phrases and binary reordering operations. In Figure 3 the sentence is decomposed, with phrases of different granularities being captured. The *REG* and *INV* tags illustrate the reordering capacities of the ITG trees, with *REG* being a monotone alignment ordering and *INV* flipping the English side with respect to the foreign phonemes. The advantage of this joint learning approach over GIZA++, Model 3P and UWS GIZA++ is that the segmentation on the phoneme side can be informed by the English, which has been shown to be valuable [20, 21, 22]. Furthermore, the base distribution Bayes ITG draws from uses cooccurrence probabilities of phrases. This contrasts with Model 3P's initialization, which uses only the limited phoneme–word alignments of GIZA++.

3. Experimental Setup

3.1. Data

To train the translation models we used the German–English parallel corpus from Europarl v7 [29]. In order to imitate a phoneme transcription, we converted the German side to a sequence of phonemes (represented with the SAMPA[5] phoneme alphabet) using the MARY text-to-speech system [30]. For example, ‘dieser’ is represented as a sequence of space-separated phonemes, ‘d i: z 6’.

The phonemic output of MARY includes some information that cannot reasonably be detected by an ASR system. In particular, stress markers and syllable boundaries are features output by the system (‘´’ and ‘-’ respectively), so we filtered them out. The granularity of tokens on the source side was thus at the phoneme level while English words were used on the target side.

Small quantities of data were used in order to mimic the realities of data collection for endangered languages. We experimented with varying data sizes to evaluate how the best method's performance scales. We used data sets of 1k, 2k, 5k, and 10k parallel sentences (corresponding to between ~13k and ~132k words), a quantity that is vastly smaller than what is typically used in statistical machine translation experiments but which approaches reasonable size for reliable manual transcription. We limited training sentences to those fewer than 100 phonemes in length.

3.2. Translation Model Training Parameters

GIZA++ was trained using the train-model.perl script included in Moses with default settings, using the grow-diag-final-and heuristic for symmetrization/phrase extraction and the msd-bidirectional-fe reordering model.

PISA was trained with default settings.

UWS GIZA++ was trained by running pgibbs first, and then running GIZA++ over the segmented phoneme sequences with default settings. The pgibbs settings were default, with the following exceptions: block sampling was used with a block size of 50, a Pitman-Yor distribution was used, and 1000 iterations were run. The final sample output by pgibbs was used as input to GIZA++. GIZA++ was run in the same way as above, using train-model.perl with heuristics for phrase extraction. It's worth noting that the hyperparameters supplied to pgibbs dictate segmentation granularity. Were they to change, we would expect the average length of the word units learnt to be different.

We ran pialign for 10 iterations with the base distribution being a log-linear interpolation of phrase cooccurrence probabilities in both directions (with a discount of 5), a beam width of 10^{-6} and a batch length of 40. The final sample was used for the purposes of phrase table extraction.

3.3. Bilingual Lexicon Extraction

To create bilingual lexicons using the above approaches, entries in the phrase tables were first sorted according to their joint probabilities. We only included entries where the length of the phonemic side was 2 or greater. This heuristic was used since it removed many spurious entries where one foreign phoneme was aligned to an entire word. Additionally, for a given English entry no more than the top 5 translations were included. A similar filter was applied to prevent more than 5 English translations of a given phoneme sequence. The top 500 entries of each lexicon were then manually annotated.

3.4. Annotation

Entries in the lexicon were evaluated by a native German speaker.[6] They were determined to be correct, incorrect or ambiguous. Correct entries are those that can readily be found in existing German–English dictionaries. For example, the entry *vIs@n⇔know* (‘wissen’). Incorrect entries are those whose translations are deemed to be clearly incorrect

[5] http://www.phon.ucl.ac.uk/home/sampa/german.htm

[6] We measured inter-annotator agreement by doubly annotating a sample of 1k entries, using a non-native German speaker, resulting in $\kappa = 0.69$.

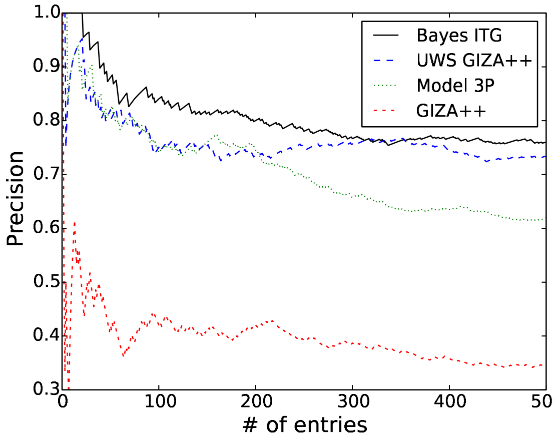

Figure 4: Comparison of the methods' precisions over the 10k dataset. Note that these are the results of the strict evaluation.

by the annotator. These include entries such as *tsu:?aIn⇔the* and *b@dINUN⇔be*. In the latter case, note that although the word alignment is incorrect, the phonemes represent a correctly segmented German word, 'Bedingung'.

Ambiguous entries are those that are neither strictly correct nor incorrect. These include entries that have boundary errors. For example, *nvi:6⇔we* ('wir') includes an extra 'n' in an otherwise correct entry. Other ambiguous entries are those that, while not found in lexicons, are nonetheless meaningful. These usually highlight interesting linguistic phenomena. For example, *nICt⇔does not* ('nicht') couldn't be found in Leo,[7] however it captures a meaningful grammatical relationship between the languages. Consider the phrase 'er rennt nicht' and one English translation 'he does not run', where this entry makes sense.

4. Quantitative Evaluation

4.1. Precision at k over bilingual entries

We compare the four models described in Section 2, each of which takes as input sentences of unsegmented phonemes and English translations. Figure 4 shows the precisions of the bilingual lexicons as the number of entries increases from 1 to 500 (sorted by the joint probability given by the model), using the methods trained on 10k sentences.[8] The 'traditional' approach with GIZA++ is the worst performer across the board. This is to be expected as it uses lexical translation probabilities between poorly translated German phonemes and English words as the basis for the extracted phrases. As a point of comparison to these models, we trained an 'oracle' model on correctly segmented phonemes using GIZA++,

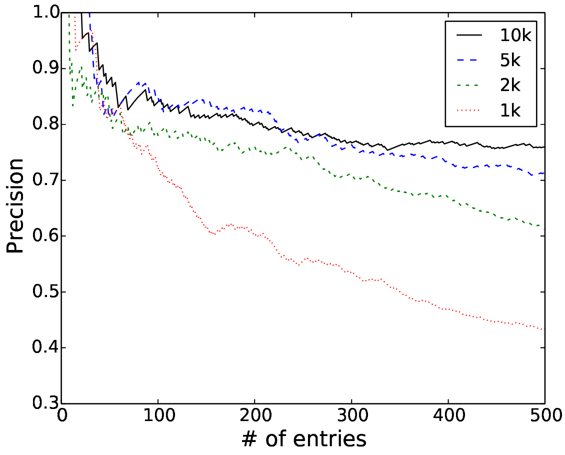

Figure 5: Comparison of Bayes ITG precisions for different sized data.

which removes the effect of segmentation errors but still includes the effect of alignment errors. This oracle model yielded a lexicon with a precision of 0.932 over the top 500 entries.

The other methods are more similar in performance, with the best performing approach being Bayes ITG. Though the results are close, the better performance of Bayes ITG as compared to the unsupervised word segmentation approach can possibly be attributed to the added information the English side provides in determining useful German phrases. This contrasts to the unsupervised word segmentation approach which segments using only monolingual German phonemic data. Performance gains over PISA's Model 3P can perhaps be attributed to limitations in Model 3P's generative model. Rather than learning explicit phrasal relationships between phoneme groups and words, Model 3P conditions the generation of phonemes from latent words and the location within that word.

Similar trends in the scores were demonstrated when evaluating precisions that accepted ambiguous entries as also correct.

Given that Bayes ITG was the best-performing approach on 10k sentences, we additionally evaluated it on smaller data sizes (see Figure 5). The fewer sentences of phonemes that are supplied the more reasonable it is to assume that they can be acquired through reliable manual transcription in a real language preservation scenario. Precision appears to be a logarithmic function of the size of the training data. These results suggest that the first few hundred entries in a lexicon can be acquired with good precision even with very limited data.

4.2. Word segmentation performance

In addition to evaluating the quality of the bilingual entries, we evaluated the quality of monolingual lexical entries on

[7] http://www.leo.org

[8] Note that we do not investigate recall as it is both difficult to establish and less relevant in the early stages of language documentation as only a small fraction of words will be captured in any case.

Method	Sents	Incorrect %	Correct seg. %
Bayes ITG	1k	26.2	52.7
Bayes ITG	2k	16.6	60.2
Bayes ITG	5k	13.4	62.7
Bayes ITG	10k	9.6	62.5
UWS GIZA++	10k	7.2	38.9
GIZA++	10k	19.4	15.5
Model 3P	10k	14.6	46.6

Table 1: The accuracy of the segmentation of phonemic lexical entries judged incorrect. The *Incorrect %* columns indicate the percentage of the 500 annotated entries that were labeled completely *incorrect* as bilingual entries . The *Correct seg. %* column indicates the percentage of those *incorrect* entries that were correctly segmented monolingual entries.

the phoneme side. This is motivated by the observation that often correct phonemic word units were extracted but mistranslated. Since monolingual entries are useful in their own right for language documentation purposes (for instance, as a useful starting point for manual correction) and language modeling, we assessed entries that were *incorrect* to determine whether the phonemic component was segmented correctly at the word boundaries.

Table 1 shows the proportion of the total entries that were annotated as *incorrect* and the proportion of those entries that were correct monolingual lexical entries on the phoneme side. Bayes ITG demonstrates effective inference of lexical items with few boundary errors, outperforming the other methods regardless of the amount of training data used. This corroborates past research that indicates that word segmentation can be better informed with bilingual data [20, 21, 22].

Also noteworthy is the outperformance of Model 3P relative to UWS GIZA++ when entries are *correct* (though having fewer strictly *incorrect* entries overall). In the approach of UWS GIZA++ it is impossible to break apart phoneme groups that have been chunked across word boundaries by the monolingual segmentation phase. However, the other methods aren't constrained by early, poorly informed chunking. This allows Model 3P relatively better word segmentation despite lower precision of bilingual lexical entries.

Note that although we are evaluating monolingual entries, the entries of UWS GIZA++ are still informed by the alignments with English, as the entries evaluated are the highest probability bilingual lexical entries found. This mitigates the problem of the effort required to tweak the hyperparameters of the word segmenter to find the right granularity of phoneme clusters. The granularity is instead informed by the English. To appreciate this, consider the most occurring lexical entries of the monolingual supervision *without* being informed by the alignments, as shown in Table 2. Of these, the only one that is an actual word is *di:* ('die'). The rest are common sub-word units. Note though that *@n* ('-en') is a common suffix for infinitive verbs—a particularly useful morpheme.

Token	Occurrences
?	13,096
@	8,587
n	8,138
t	6,422
@n	6,300
d	5,929
s	3,226
6	3,136
f	3,099
di:	2,913

Table 2: The most common lexical entries found by the unsupervised word segmentation, without harnessing bilingual information.

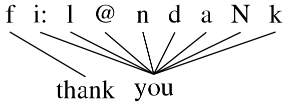

Figure 6: The phonemes of *vielen dank* as aligned to *thank you* by PISA's Model 3P.

5. Qualitative Evaluation

To appreciate the peculiarities and differences of these approaches, we will now consider some general observations made by examining the lexicons of the various approaches, discussing some representative lexical entries and word alignments.

Model 3P seemed generally more susceptible to off-by-one errors at the boundaries of entries. A high confidence, but incorrect, entry that occurred in the lexicon based on Model 3P alignments was *i:l@ndaNk⇔you* ('vielen dank'). The English makes some sense, as *vielen dank* can be translated as 'thank you' or 'thank you very much', although the 'thank' component on the English side is missing. Notably, the German side is segmented incorrectly at the phrase boundary, missing the initial phoneme 'f' (it should be 'fi:l@ndaNk'). It turns out that in sentences containing this German phoneme sequence, the 'f' is often aligned to English 'thank' (see Figure 6). In the lexicons created by both Bayes ITG and UWS GIZA++ this entry was correctly phrase-segmented as 'fi:l@ndaNk'.

A similar such entry in the Model 3P lexicon was *daspa6la:mEn⇔parliament*, where the source side is missing the final 't'. In the lexicon constructed using Bayes ITG, such boundary mistakes were scarce. The equivalent entry was *daspa6la:mEnt⇔parliament* ('das Parlament'). Note that this entry was not considered strictly correct nor correctly segmented, as it is comprised of two words, with the German article being included. However in this case, as in almost all others, Bayes ITG still segments correctly at the boundaries of multiword units (as distinct from correctly segmented individual words). One of the instances of an entry annotated as *incorrect* in the top 500 entries of the Bayes ITG lexicon where the phoneme side was also incorrectly segmented was *tvo6d@n⇔been*, where there is a spurious '*t*' prefixing the phonemic representation of 'worden'. Investigating the alignments highlights the cause of this entry. Phoneme sequences such as *Unt6StYtstvO6d@n* ('unterstützt worden') and ?E6RaICtvO6d@n ('erreicht worden') include verbs that often appear inflected with different suffixes elsewhere, but end in 't' when occurring before *vO6d@n* ('unterstützen' and 'erreichen' respectively, with the suffix '-en'). High correlation of *vO6d@n* ('worden') and the suffix '*t*' likely caused this entry.

The lexicon constructed using Model 3P demonstrated an apparent bias to shorter units. In that lexicon, the above entry was segmented correctly as *vO6d@n*. On the other hand, Bayes ITG tended to learn towards longer multiword units, as a result of the model's capacity to capture phrases at coarser granularities. *po:Ete:6RIN⇔Mr Poettering* was present in the Bayes ITG lexicon, but not in the others. The title is missing on the source side. This can be attributed to varying morphology of the title, which takes the form of both 'Herr' and 'Herrn' depending on context. However, since the English side consistently takes the form of 'Mr Poettering', evidence is built up primarily to relate both the title and name on the English side to only the name on the phoneme side.

For all the alignment approaches, there were many entries that are justified given only the information present in the corpus. The above example, *daspa6la:mEn⇔parliament*, is one such example and is arguably correct in some contexts (consider the phrase *das Parlament lehnte den Antrag ab⇔Parliament rejected the request*). This entry can be attributed to linguistic differences that possibly no alignment algorithm can overcome, with the article often being optional in English translations. In general, the entries Bayes ITG presented us with tend to be interpretable with respect to how phoneme sequences occur in the corpus.

UWS GIZA++ yielded the high confidence, yet erroneous, entries *t?⇔is*, *n?⇔to*, *n?⇔of*, which didn't occur in the other lexicons. This is likely a result of the pipelined nature of the approach, where monolingual segmentation is first performed before alignment. The German components to these entries represent frequently occurring phonemic sequences (many words end with '*t*' or '*n*' and many start with a glottal stop, '*?*', before some vowel). The English sides represent function words that are so commonly occurring that the coincidental cooccurrence of these phonemes and English words allowed them to become extracted lexical entries, which were not obtained using Bayes ITG or Model 3P. Entries such as this partly explain why UWS GIZA++ failed to perform as well as Model 3P in segmenting lexical entries despite outperforming it in bilingual precision. The other likely reason is that chunks that cross word boundaries learnt during monolingual segmentation cannot be undone.

6. Conclusion

We compared four representative approaches, evaluating the quality of monolingual and bilingual lexical entries. While two of the techniques had been previously established for the task of phoneme–word alignment, we achieved performance improvements by applying models that had not previously been considered for this task, demonstrating that hundreds of bilingual lexicon entries can be learnt with as few as 1k sentences of bilingual data. This can be done despite using an unsegmented phonemic representation of the source side.

Such approaches may be used to indicate what can be inferred from corpora of interleaved audio in the absence of reliable segmentation, aid in post-mortem linguistic analysis of a language, and to bootstrap ASR systems in order to help improve their phoneme recognition.

7. References

[1] T. Hughes, K. Nakajima, L. Ha, A. Vasu, P. J. Moreno, and M. LeBeau, "Building transcribed speech corpora quickly and cheaply for many languages." in *INTERSPEECH*, 2010, pp. 1914–1917.

[2] 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.

[3] N. J. De Vries, M. H. Davel, J. Badenhorst, W. D. Basson, F. De Wet, E. Barnard, and A. De Waal, "A smartphone-based ASR data collection tool for underresourced languages," *Speech communication*, vol. 56, pp. 119–131, 2014.

[4] D. W. Reiman, "Basic oral language documentation," in *Language Documentation & Conservation*. University of Hawai'i Press, December 2010, pp. 254–268.

[5] 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, June 2014, pp. 1–5.

[6] D. Wu and X. Xia, "Learning an English-Chinese lexicon from a parallel corpus," in *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, 1994, pp. 206–213.

[7] I. D. Melamed, "Automatic construction of clean broad-coverage translation lexicons," *CoRR*, 1996.

[8] H. M. Caseli, V. N. Maria das Graças, and M. L. Forcada, "Automatic induction of bilingual resources from aligned parallel corpora: application to shallow-transfer machine translation," *Machine Translation*, vol. 20, no. 4, pp. 227–245, 2006.

[9] A. Lardilleux, J. Gosme, and Y. Lepage, “Bilingual lexicon induction: Effortless evaluation of word alignment tools and production of resources for improbable language pairs,” in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC-2010)*. Valletta, Malta: European Languages Resources Association (ELRA), May 2010, pp. 252–256.

[10] S. Stüker and A. Waibel, “Towards human translations guided language discovery for ASR systems.” in *SLTU*, 2008, pp. 76–79.

[11] S. Stüker, L. Besacier, and A. Waibel, “Human translations guided language discovery for ASR systems.” in *INTERSPEECH*, 2009, pp. 3023–3026.

[12] P. E. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer, “The mathematics of statistical machine translation: Parameter estimation,” *Computational Linguistics*, vol. 19, no. 2, pp. 263–312, 1993.

[13] F. Stahlberg, T. Schlippe, S. Vogel, and T. Schultz, “Word segmentation through cross-lingual word-to-phoneme alignment,” in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 85–90.

[14] ——, “Pronunciation extraction from phoneme sequences through cross-lingual word-to-phoneme alignment,” in *Statistical Language and Speech Processing*. Springer, 2013, pp. 260–272.

[15] ——, “Towards automatic speech recognition without pronunciation dictionary, transcribed speech and text resources in the target language using cross-lingual word-to-phoneme alignment,” in *Spoken Language Technologies for Under-Resourced Languages*, 2014.

[16] ——, “Word segmentation and pronunciation extraction from phoneme sequences through cross-lingual word-to-phoneme alignment,” *Computer Speech & Language*, pp. –, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885230814000977

[17] C. Cieri and M. Liberman, “More data and tools for more languages and research areas: a progress report on ldc activities,” in *5th International Conference on Language Resources and Evaluation (LREC2006), Genoa, Italy*, 2006.

[18] S. Bird and D. Chiang, “Machine translation for language preservation.” in *24th International Conference on Computational Linguistics*, 2012, p. 125.

[19] Y. Deng and W. Byrne, “HMM word and phrase alignment for statistical machine translation,” in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada: ACL, October 2005, pp. 169–176.

[20] J. Xu, J. Gao, K. Toutanova, and H. Ney, “Bayesian semi-supervised Chinese word segmentation for statistical machine translation,” in *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. ACL, 2008, pp. 1017–1024.

[21] P.-C. Chang, M. Galley, and C. D. Manning, “Optimizing Chinese word segmentation for machine translation performance,” in *Proceedings of the Third Workshop on Statistical Machine Translation*. Columbus, Ohio: ACL, June 2008, pp. 224–232.

[22] T. Nguyen, S. Vogel, and N. A. Smith, “Nonparametric word segmentation for machine translation,” in *Proceedings of the 23rd International Conference on Computational Linguistics*. ACL, 2010, pp. 815–823.

[23] F. J. Och and H. Ney, “A systematic comparison of various statistical alignment models,” *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[24] D. Mochihashi, T. Yamada, and N. Ueda, “Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling,” in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Suntec, Singapore: ACL, August 2009, pp. 100–108.

[25] G. Neubig, “Simple, correct parallelization for blocked Gibbs sampling,” Nara Institute of Science and Technology, Tech. Rep., 2014. [Online]. Available: http://www.phontron.com/paper/neubig14pgibbs.pdf

[26] G. Neubig, T. Watanabe, S. Mori, and T. Kawahara, “Machine translation without words through substring alignment,” in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Jeju Island, Korea: ACL, July 2012, pp. 165–174.

[27] G. Neubig, T. Watanabe, E. Sumita, S. Mori, and T. Kawahara, “An unsupervised model for joint phrase alignment and extraction,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 632–641.

[28] D. Wu, “Stochastic inversion transduction grammars and bilingual parsing of parallel corpora,” *Computational linguistics*, vol. 23, no. 3, pp. 377–403, 1997.

[29] P. Koehn, “Europarl: A parallel corpus for statistical machine translation,” in *MT Summit*, vol. 5, 2005, pp. 79–86.

[30] M. Schröder and J. Trouvain, "The German text-to-speech synthesis system MARY: A tool for research, development and teaching," *International Journal of Speech Technology*, vol. 6, no. 4, pp. 365–377, 2003.