Spoken Term Discovery for Language Documentation using Translations

Antonios Anastasopoulos[♠*] **Sameer Bansal**[◇*]
Sharon Goldwater[◇] **Adam Lopez**[◇] **David Chiang**[♠]
[♠]Department of Computer Science and Engineering, University of Notre Dame
[◇]School of Informatics, University of Edinburgh

Abstract

Vast amounts of speech data collected for language documentation and research remain untranscribed and unsearchable, but often a small amount of speech may have text translations available. We present a method for partially labeling additional speech with translations in this scenario. We modify an unsupervised speech-to-translation alignment model and obtain prototype speech segments that match the translation words, which are in turn used to discover terms in the unlabelled data. We evaluate our method on a Spanish-English speech translation corpus and on two corpora of endangered languages, Arapaho and Ainu, demonstrating its appropriateness and applicability in an actual very-low-resource scenario.

1 Introduction

Language documentation efforts over the last 50–60 years have resulted in audio recordings of native speakers in a large number of languages, many of which are available online. However, due to the enormous effort required for transcription, much of the data remains unannotated and unsearchable.[1] For example, out of the 137 unrestricted collections in the Archive of the Indigenous Languages of Latin America, about half (49%) contain no transcriptions at all, and only 7% are fully transcribed.[2] As a result, some recent documentation efforts have begun to focus instead on annotating with *translations*, often with the help of bilingual native speakers themselves (Bird et al., 2014; Blachon et al., 2016; Adda et al., 2016).

Nevertheless, even translation takes time and language knowledge, so there may still be little translated data relative to the amount of recorded audio. An important goal, then, is to bootstrap language technology from this small parallel corpus in order to provide tools to annotate more data or make the data more searchable.

We build on the approach of Anastasopoulos et al. (2016), who developed a system that performs joint inference to identify recurring segments of audio and cluster them while aligning them to words in a text translation. Here, we extend the method to be able to search for new instances of the latent clusters within the unlabeled audio, effectively providing keyword translations for some of the unlabeled speech. We evaluate our method on a Spanish-English corpus used in previous work, and on two datasets from endangered languages (narratives in Arapaho and Ainu). No previous computational methods have been tested on the latter data, to our knowledge. We show that in all cases, our system outperforms a recent baseline targeted at the same very low-resource setting (Bansal et al., 2017b), also showing robustness to audio quality and preprocessing decisions.

2 Related work

Our work joins a handful of other recent proposals aimed at low-resource speech-to-text alignment and translation. These include those of Duong et al. (2016) and Anastasopoulos et al. (2016), who performed alignment only; Bérard et al. (2016), who used synthetic rather than real speech; and Adams et al. (2016) and Godard et al. (2016), who worked from phone lattices and phone sequences, respectively; Stahlberg et al. (2013), who perform phone-to-translation alignment for pronunci-

∗ Equal contribution.

[1]By some estimates, a trained linguist requires up to one hour for to phonetically transcribe one minute of speech (Thi-Ngoc-Diep Do and Castelli, 2014).

[2]http://ailla.utexas.org

ation extraction. Weiss et al. (2017) presented a sequence-to-sequence neural model that learned a direct mapping from speech to translated text with impressive results, but was trained on roughly 140 hours of parallel data—far more than is available for most endangered languages.

The only previous system we know of to address the same very-low-resource scenario and provide translation terms for unlabeled audio is that of Bansal et al. (2017b) (henceforth `UTD-align`), who used an unsupervised term discovery system (Jansen et al., 2010) to cluster recurring audio segments into pseudowords. The pseudowords occurring in the parallel section of the corpus were then aligned to the translation text using IBM Model 1, and used to translate instances occurring in the test (audio-only) section.

3 Method

The main difference between our method and `UTD-align` is that `UTD-align` clusters the audio prior to aligning with the translations, whereas we start by performing joint alignment and clustering using an improved version of the method proposed by Anastasopoulos et al. (2016) (henceforth `s2t`). The resulting aligned clusters are represented by one or more prototype speech segments. We extend `s2t` to identify new instances of those prototypes in the unlabeled speech, using a modified version of `ZRTools`, the same UTD toolkit used by `UTD-align`.[3] (Jansen et al., 2010)

Previous work has indicated that using translation text to inform acoustic clustering provides more accurate clusters than just using UTD (Bansal et al., 2017a), so we initially expected that this straightforward extension of `s2t` would work better than `UTD-align`. However, early experiments indicated that the text had *too* much influence on clustering, yielding clusters with highly diverse audio, and thus poor prototypes. Thus, we modified `s2t`[4] in order to account for this issue, obtaining prototypes of higher quality (§3.1), which we search for in the unlabeled audio (§3.2).

3.1 Aligning speech to translation

The `s2t` model is an extension of IBM Model 2 for word alignment (Brown et al., 1993), combined with K-means clustering using Dynamic Time Warping (DTW) (Berndt and Clifford, 1994) as a distance measure. It uses expectation-maximization (EM) to align speech segments to words in the parallel text, while jointly clustering the segments. Each translation word is aligned to an acoustic segment, with overlapping alignments and unaligned speech spans being allowed.

In the original implementation, every translation word was represented by a fixed number (2) of acoustic sub-clusters, with a single prototype representing each.[5] The prototypes are averages of the segments in the cluster, computed using DTW Barycenter Averaging (Petitjean et al., 2011). At the E-step, each segment was assigned to its closest sub-cluster, and at the M-step the sub-cluster's prototype was re-computed. However, the original choice of two subclusters was fairly arbitrary, and we found it doesn't sufficiently account for the wide acoustic variability due to gender or speaker. We thus modify `s2t` so that, before the M-step, each cluster's segments are grouped into sub-clusters using connected components clustering with a similarity threshold δ, following Park and Glass (2008). That way, the number of sub-clusters and prototypes for each translation word is determined automatically based on the acoustic similarity of the segments.

Our preliminary analysis showed that shorter alignments tend to introduce significantly more noise than longer ones. Therefore, in the final M-step of `s2t`, we discard all segments shorter than a length threshold t before computing the prototypes. We use the default values for the rest of the `s2t` parameters.

Another pragmatic choice we made based on the performance of our method was to remove the stopwords from the translations, following Bansal et al. (2017b). The rationale is that translation stopwords would not be particularly useful for labelling speech in our envisioned use cases.

3.2 Keyword Search

In the second stage, we use the approximate DTW-based pattern matching method of `ZRTools` to search for the obtained prototypes in the test data. We require that each discovered term matches at least $k\%$ of a prototype's length and that its DTW similarity score is higher than a threshold s. By varying s we can control the number of discovered terms, trading off precision and recall. Also, we do not allow overlapping matches; in the case

[3] https://github.com/arenjansen/ZRTools

[4] The code is available at https://bitbucket.org/ndnlp/translationTermDiscovery

[5] https://bitbucket.org/ndnlp/speech2translation

of an overlap, we output the match with the higher score.

4 Experiments

The CALLHOME Spanish Speech dataset (LDC2014T23) with English translations (Post et al., 2013) has been used in almost all groundlaying previous work, treating Spanish as a low-resource language. As a collection of telephone conversations between relatives (about 20 total hours of audio), it doesn't match our language documentation scenario, but we use it in order to compare our method with previous work.

We shuffle the utterances and split them into training, dev, and test sets with 70%, 10%, and 20% of the data, respectively. We filter the active audio regions using energy-based voice activity detection (VAD). We obtain prototypes in the training set and tune the values of the length threshold t, the similarity threshold d, and the partial overlap threshold k on the development set using grid search. The best parameter combination is $t = 300$ ms, $d = 90\%$, and $k = 80\%$, while $s = 0.90$ returns the highest F-score. We evaluate our discovered translation terms on the test set using precision, recall, and F-score at the token level over the correct bag-of-words translations.

We also evaluate our method on two low-resource endangered languages, Arapaho and Ainu. For these experiments, we only have a training and test set, so we use the same preprocessing and hyperparameter settings as in CALLHOME.

Arapaho is an Algonquian language with about 1,000 native speakers, mostly in Wyoming. We use 8 narratives published at The Arapaho Language Project,[6] which provides the narratives' audio along with English translations, among other language learning resources.

Hokkaido Ainu is the sole surviving member of the Ainu language family and is generally considered a language isolate. As of 2007, only ten native speakers were alive. The Glossed Audio Corpus of Ainu Folklore provides 10 narratives with audio and translations in English.[7] More information and statistics on the Arapaho and Ainu corpora is provided in Tables 4 and 5.

[6] http://www.colorado.edu/csilw/alp/index.html
[7] http://ainucorpus.ninjal.ac.jp/corpus/en/

Method	Prec	Rec	F-score	Coverage
`UTD-align`	5.1	2.1	3.0	27%
ours	4.2	3.5	3.8	59%
ours (oracle)	5.3	4.9	5.1	65%

Table 1: Results of our method and baseline work on the CALLHOME dataset. Our method improves over `UTD-align` whether inferring alignments or using oracle (silver) alignments.

4.1 Results on CALLHOME

We first evaluate the effect of our modifications to the `s2t` method, by calculating alignment F-score on links between speech frames and translation words.[8] The intermediate sub-clustering step between the E- and M-steps results in a more informed selection of the number of sub-clusters that increases the alignment F-score by 1.5%. Also, removing translation stopwords further leads to higher alignment precision by +4%. Alignment recall is lower since it's computed over the alignments of both content and stopwords. Although both improvements are small, the higher alignment precision leads to better prototypes.

In addition, Duong et al. (2016) created "silver" standard speech-to-translation alignments by combining the forced speech-to-transcription alignments and the transcription-to-translation word alignments. These are useful for evaluating how well the prototype creation and matching could work, given oracle speech-to-translation alignments. In Table 1, we report precision, recall, and F-score on the discovered translation terms (at the token level) using prototypes from both "silver" and noisy alignments. We also report the percentage of active audio that is labelled (coverage). In both cases we outperform `UTD-align`.[9] Even though there is room for improvement, using the translation information at the alignment stage certainly improves the clustering, as anticipated. Another advantage of our method over `UTD-align` is its significantly improved coverage of the active audio, as shown in the last column of Table 1. The precision-recall curve obtained by varying the output similarity threshold s is shown in Figure 1.

[8] See the paper by Duong et al. (2016) for a full definition.
[9] The code was provided by the authors of `UTD-align`.

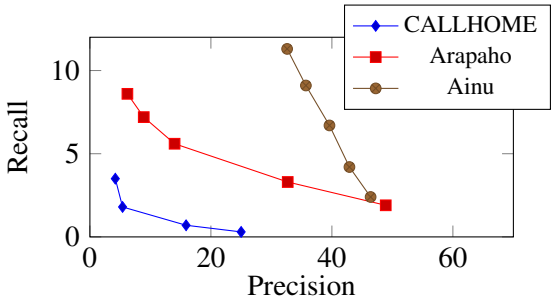

Figure 1: Average precision and recall curve for our discovered matches in CALLHOME and the Arapaho and Ainu test narratives (varying the output threshold s between 0.90 and 0.94).

Arapaho narrative	Terms found	Prec (%)	Rec (%)	Oracle Recall
1	29	31.0	4.7	32.3
2	65	21.5	8.0	44.3
3	91	7.7	6.4	54.5
4	158	13.9	8.4	53.4
6	1	100.0	0.7	41.4
7	104	7.7	7.1	44.6
8	10	30.0	4.5	65.2
average-ours	65	14.0	6.0	
`UTD-align`	2	26.7	0.4	

Table 2: Results on Arapaho narratives. In general, we identify meaningful translation terms.

4.2 Results on Arapaho and Ainu

Out of the eight Arapaho narratives, we select the longest (18 minutes of audio, 233 English word types) for training, using the other seven (32 minutes total) for evaluation. The Ainu collection provides ten narratives, so we use the first two for training (24 minutes of audio, 494 English word types) and the rest (133 minutes total) as test data.

Treating each narrative as a bag of words, the precision and recall results at the token level are shown in Tables 2 and 3. The last columns of these Tables correspond to the highest possible recall that we could get if we discovered all the training terms that also appear in the test set. Precision-recall curves can be seen in Figure 1.

On both corpora, `UTD-align` identifies hardly any translation terms, with recall scores below 1% and average F-scores of 0.8% and 0.2% for Arapaho and Ainu, respectively. Preprocessing with the same VAD script as for our method, `UTD-align` produced too many spurious matches

Ainu narrative	Terms found	Prec (%)	Rec (%)	Oracle Recall
3	80	50.0	3.8	63.0
4	73	49.3	4.5	67.1
5	199	49.7	5.1	61.8
6	174	22.4	9.0	65.0
7	123	19.5	8.9	56.1
8	122	57.4	3.9	67.8
9	59	62.7	1.5	63.0
10	149	46.3	6.6	69.7
average-ours	122	42.3	4.2	
`UTD-align`	4	24.2	0.1	

Table 3: Results on the Ainu narratives. We are able to correctly identify several terms per story, with quite high precision.

(millions); we then used a more aggressive filtering which removed more parts of the audio, but it resulted in too few discovered matches (as shown here). In principle, it should be possible to tailor the preprocessing parameters for each corpus and improve results for `UTD-align`.

Our method, instead, outputs several terms per narrative without the need to readjust preprocessing decisions, with F-scores of 8.4% (Arapaho) and 7.2% (Ainu). Two exceptions are Arapaho narratives #6 and #8, which, unlike our training data, are narrated by a woman. Although there is clearly room for improvement in terms of recall, as shown by the last columns of Tables 2 and 3, we are generally able to identify meaningful terms.

For most of the Arapaho stories we discover named entities such as *Ghost* and *Strong Bear*, content nouns like *tipis* and *mountains*, or verbs such as *hunting*. In Ainu we discover more terms, but the narratives are also longer. A larger domain shift between training and test (small overlap on named entities and other content words) leads to lower recall compared to Arapaho. Our method correctly identifies mostly common terms in the Ainu narratives, like *village*, *food*, as well as verbs used in narration such as *said*, *went*, or *came*.

5 Conclusion

We propose a method that modifies and extends a speech-to-translation alignment method and can be used for identifying translation terms in unlabeled audio, appropriate for extremely small datasets. On CALLHOME, we show small

improvements over a recent baseline. We also demonstrate the applicability of our method on language documentation scenarios, by applying it on two endangered language datasets. Speaker differences are still an issue, but our method is more robust to differences in acoustic quality than the previous method.

Acknowledgements We are grateful for support from NSF Award 1464553. This work was also supported in part by a James S McDonnell Foundation Scholar Award and a Google faculty research award. Goldwater is the recipient of James S. McDonnell Foundation Scholar Award #220020374.

References

Oliver Adams, Graham Neubig, Trevor Cohn, Steven Bird, Quoc Truong Do, and Satoshi Nakamura. 2016. Learning a lexicon and translation model from phoneme lattices. In *Proc. EMNLP*, pages 2377–2382.

Gilles Adda, Sebastian Stüker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, et al. 2016. Breaking the unwritten language barrier: The BULB project. *Procedia Computer Science*, 81:8–14.

Antonios Anastasopoulos, David Chiang, and Long Duong. 2016. An unsupervised probability model for speech-to-translation alignment of low-resource languages. In *Proc. EMNLP*, pages 1255–1263.

Sameer Bansal, Herman Kamper, Sharon Goldwater, and Adam Lopez. 2017a. Weakly supervised spoken term discovery using cross-lingual side information. In *Proc. ICASSP*.

Sameer Bansal, Herman Kamper, Adam Lopez, and Sharon Goldwater. 2017b. Towards speech-to-text translation without speech recognition. In *Proc. EACL, Vol. 2*, pages 474–479.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *Proc. NIPS End-to-end Learning for Speech and Audio Processing Workshop*.

Donald J. Berndt and James Clifford. 1994. Using dynamic time warping to find patterns in time series. In *Proc. KDD*, pages 359–370.

Steven Bird, Lauren Gawne, Katie Gelbart, and Isaac McAlister. 2014. Collecting bilingual audio in remote indigenous communities. In *Proc. COLING*, pages 1015–1024.

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, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. 2016. An attentional model for speech translation without transcription. In *Proc. NAACL-HLT*, pages 949–959.

Pierre Godard, Gilles Adda, Martine Adda-Decker, Alexandre Allauzen, Laurent Besacier, Helene Bonneau-Maynard, Guy-Noël Kouarata, Kevin Löser, Annie Rialland, and François Yvon. 2016. Preliminary experiments on unsupervised word discovery in Mboshi. In *Proc. INTERSPEECH*.

Aren Jansen, Kenneth Church, and Hynek Hermansky. 2010. Towards spoken term discovery at scale with zero resources. In *Proc. INTERSPEECH*, pages 1676–1679.

Alex S. Park and James R. Glass. 2008. Unsupervised pattern discovery in speech. *IEEE Trans. Audio, Speech, and Language Processing*, 16(1):186–197.

François Petitjean, Alain Ketterlin, and Pierre Gançarski. 2011. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3):678–693.

Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. 2013. Improved speech-to-text translation with the fisher and callhome spanish-english speech translation corpus. In *Proc. IWSLT*.

Felix Stahlberg, Tim Schlippe, Stephan Vogel, and Tanja Schultz. 2013. Pronunciation extraction from phoneme sequences through cross-lingual word-to-phoneme alignment. In *ICSLSP*, pages 260–272. Springer.

Alexis Michaud Thi-Ngoc-Diep Do 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 *Proc. SLTU*, pages 153–160.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly transcribe foreign speech. arXiv:1703.08581.

ID	Title	Duration (m:s)	Transcription Tokens	Transcription Types	Translation Tokens	Translation Types
1	Fooling the ghost	5:12	134	91	192	80
2	The Ghost by the Road	7:00	140	104	176	117
3	The Old Couple and the Ghost	3:12	88	71	110	74
4	The Owl Man	7:14	269	157	262	125
5	Strong Bear and the Ghost	18:35	523	346	591	289
6	The Woman who turned into Stone	3:26	140	93	152	85
7	Strong Bear and the Boxer	3:29	125	82	112	61
8	Telescope	1:40	54	48	66	48
	total	50:00	1473	849	1661	556

Table 4: Statistics on the Arapaho narratives. English type and token counts do not include stopwords.

ID	Title	Duration (m:s)	Transcription Types	Transcription Tokens	Translation Types	Translation Tokens
1	Pananpe escapes from the demons hands	6:12	189	849	203	519
2	The Girl who Gave the Bad Red Dog Poison	17:48	488	2634	537	1336
3	The Young Lad Raised by the Cat God	15:14	450	2149	437	1066
4	The Poor Man who Dug Up the Village Chief Wife's Grave	10:38	306	1551	365	796
5	The Grapevines which Warded Off the Topattumi-night Raiders	24:41	572	3600	660	1942
6	The Woman who Became Kemkacikappo Bird	8:59	233	699	219	431
7	The Goddess of the Fire Fought with the Demon God From the End of the Earth	6:03	161	416	156	271
8	The Bridge of Mist	23:09	519	3408	591	1816
9	The Rich Man from Cen-pak	32:59	699	4845	789	2523
10	Godly Elder Sister Gets Rid of Bad Bear Father	12:16	400	1789	401	1043
	total	157:59	1826	21940	1861	11743

Table 5: Statistics on the Ainu narratives. English type and token counts do not include stopwords.