Towards speech-to-text translation without speech recognition

Sameer Bansal[1], *Herman Kamper*[2], *Adam Lopez*[1], *Sharon Goldwater*[1]
[1]School of Informatics, University of Edinburgh
[2]Toyota Technological Institute at Chicago, USA
{sameer.bansal, sgwater, alopez}@inf.ed.ac.uk, kamperh@gmail.com

Abstract

We explore the problem of translating speech to text in low-resource scenarios where neither automatic speech recognition (ASR) nor machine translation (MT) are available, but we have training data in the form of audio paired with text translations. We present the first system for this problem applied to a realistic multi-speaker dataset, the CALLHOME Spanish-English speech translation corpus. Our approach uses unsupervised term discovery (UTD) to cluster repeated patterns in the audio, creating a *pseudotext*, which we pair with translations to create a parallel text and train a simple bag-of-words MT model. We identify the challenges faced by the system, finding that the difficulty of cross-speaker UTD results in low recall, but that our system is still able to correctly translate some content words in test data.

1 Introduction

Typical speech-to-text translation systems pipeline automatic speech recognition (ASR) and machine translation (MT) (Waibel and Fugen, 2008). But high-quality ASR requires hundreds of hours of transcribed audio, while high-quality MT requires millions of words of parallel text—resources available for only a tiny fraction of the world's estimated 7,000 languages (Besacier et al., 2014). Nevertheless, there are important low-resource settings in which even limited speech translation would be of immense value: documentation of endangered languages, which often have no writing system (Besacier et al., 2006; Martin et al., 2015); and crisis response, for which text applications have proven useful (Munro, 2010), but only help literate populations. In these settings, target translations may be available. For example, ad hoc translations may be collected in support of relief operations. Can we do anything at all with this data?

In this exploratory study, we present a speech-to-text translation system that learns directly from source audio and target text pairs, and does not require intermediate ASR or MT. Our work complements several lines of related recent work. For example, Duong et al. (2016) and Anastasopoulos et al. (2016) presented models that align audio to translated text, but neither used these models to try to translate new utterances (in fact, the latter model cannot make such predictions). Berard et al. (2016) did develop a direct speech to translation system, but presented results only on a corpus of synthetic audio with a small number of speakers. Finally, Adams et al. (2016a; 2016b) targeted the same low-resource speech-to-translation task, but instead of working with audio, they started from word or phoneme lattices. In principle these could be produced in an unsupervised or minimally-supervised way, but in practice they used supervised ASR/phone recognition. Additionally, their evaluation focused on phone error rate rather than translation. In contrast to these approaches, our method can make translation predictions for audio input not seen during training, and we evaluate it on real multi-speaker speech data.

Our simple system (§2) builds on unsupervised speech processing (Versteegh et al., 2015; Lee et al., 2015; Kamper et al., 2016b), and in particular on *unsupervised term discovery* (UTD), which creates hard clusters of repeated word-like units in raw speech (Park and Glass, 2008; Jansen and Van Durme, 2011). The clusters do not account for all of the audio, but we can use them to simulate a partial, noisy transcription, or *pseudotext*, which we pair with translations to learn a bag-of-words translation model. We test our system on the CALLHOME Spanish-English speech translation corpus (Post et al., 2013), a noisy multi-speaker corpus of telephone calls in a variety of Spanish di-

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 474–479,
Valencia, Spain, April 3-7, 2017. ©2017 Association for Computational Linguistics

alects (§3). Using the Spanish speech as the source and English text translations as the target, we identify several challenges in the use of UTD, including low coverage of audio and difficulty in cross-speaker clustering (§4). Despite these difficulties, we demonstrate that the system learns to translate some content words (§5).

2 From unsupervised term discovery to direct speech-to-text translation

For UTD we use the Zero Resource Toolkit (ZRTools; Jansen and Van Durme, 2011).[1] ZRTools uses dynamic time warping (DTW) to discover pairs of acoustically similar audio segments, and then uses graph clustering on overlapping pairs to form a hard clustering of the discovered segments. Replacing each discovered segment with its unique cluster label, or *pseudoterm*, gives us a partial, noisy transcription, or pseudotext (Fig. 1).

In creating a translation model from this data, we face a difficulty that does not arise in the parallel texts that are normally used to train translation models: the pseudotext does not represent all of the source words, since the discovered segments do not cover the full audio (Fig. 1). Hence we must not assume that our MT model can completely recover the translation of a test sentence. In these conditions, the language modeling and ordering assumptions of most MT models are unwarranted, so we instead use a simple bag-of-words translation model based only on co-occurrence: IBM Model 1 (Brown et al., 1993) with a Dirichlet prior over translation distributions, as learned by `fast_align` (Dyer et al., 2013).[2] In particular, for each pseudoterm, we learn a translation distribution over possible target words. To translate a pseudoterm in test data, we simply return its highest-probability translation (or translations, as discussed in §5).

This setup implies that in order to translate, we must apply UTD on both the training and test audio. Using additional (not only training) audio in UTD increases the likelihood of discovering more clusters. We therefore generate pseudotext for the combined audio, train the MT model on the pseudotext of the training audio, and apply it to the pseudotext of the test data. This is fair since the UTD has access to only the audio.[3]

[1] `https://github.com/arenjansen/ZRTools`

[2] We disable diagonal preference to simulate Model 1.

[3] This is the simplest approach for our proof-of-concept system. In a more realistic setup, we could use the training audio to construct a consensus representation of each pseudoterm (Petitjean et al., 2011; Anastasopoulos et al., 2016), then use DTW to identify its occurrences in test data to translate.

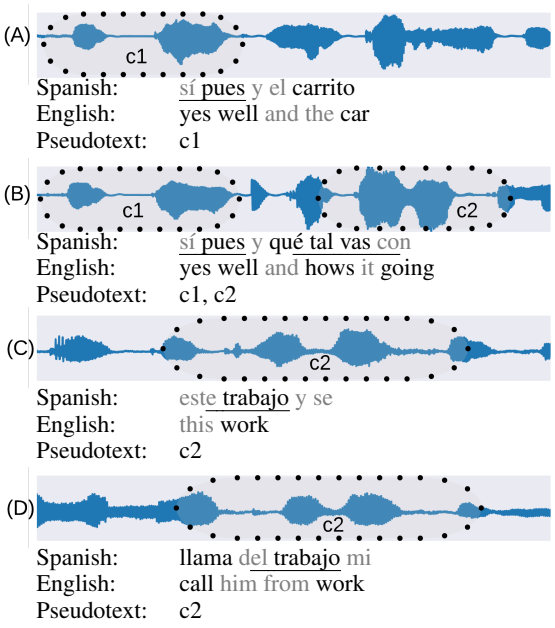

Figure 1: Example utterances from our data, showing UTD matches, corresponding pseudotext, and English translation. For clarity, we also show Spanish transcripts with the approximate alignment of each pseudoterm underlined, though these transcripts are unavailable to our system. Stopwords (in gray) are ignored in our evaluations. These examples illustrate the difficulties of UTD: it does not match the full audio, and it incorrectly clusters part of utterance B with a good pair in C and D.

3 Dataset

Although we did not have access to a low-resource dataset, there is a corpus of noisy multi-speaker speech that simulates many of the conditions we expect to find in our motivating applications: the CALLHOME Spanish–English speech translation dataset (LDC2014T23; Post el al., 2013).[4] We ran UTD over all 104 telephone calls, which pair 11 hours of audio with Spanish transcripts and their crowdsourced English translations. The transcripts contain 168,195 Spanish word tokens (10,674 types), and the translations contain 159,777 English word tokens (6,723 types). Though our system does not require Spanish transcripts, we use them to evaluate UTD and to simulate a perfect UTD system, called the *oracle*.

For MT training, we use the pseudotext and translations of 50 calls, and we filter out stopwords in the

[4] We did not use the Fisher portion of the corpus.

translations with NLTK (Bird et al., 2009).[5] Since UTD is better at matching patterns from the same speaker (§4.2), we created two types of 90/10% train/test split: at the *call level* and at the *utterance level*. For the latter, 90% of the utterances are randomly chosen for the training set (independent of which call they occur in), and the rest go in the test set. Hence at the utterance level, but not the call level, some speakers are included in both training and test data. Although the utterance-level split is optimistic, it allows us to investigate how multiple speakers affect system performance. In either case, the oracle has about 38k Spanish tokens to train on.

4 Analysis of challenges from UTD

Our system relies on the pseudotext produced by ZRTools (the only freely available UTD system we are aware of), which presents several challenges for MT. We used the default ZRTools parameters, and it might be possible to tune them to our task, but we leave this to future work.

4.1 Assigning wrong words to a cluster

Since UTD is unsupervised, the discovered clusters are noisy. Fig. 1 shows an example of an incorrect match between the acoustically similar "qué tal vas con" and "te trabajo y" in utterances B and C, leading to a common assignment to c2. Such inconsistencies in turn affect the translation distribution conditioned on c2.

Many of these errors are due to cross-speaker matches, which are known to be more challenging for UTD (Carlin et al., 2011; Kamper et al., 2015; Bansal et al., 2017). Most matches in our corpus are across calls, yet these are also the least accurate (Table 1). Within-utterance matches, which are always from the same speaker, are the most reliable, but make up the smallest proportion of the discovered pairs. Within-call matches fall in between. Overall, average cluster purity is only 34%, meaning that 66% of discovered patterns do not match the most frequent type in their cluster.

4.2 Splitting words across different clusters

Although most UTD matches are across speakers, recall of cross-speaker matches is lower than for same-speaker matches. As a result, the same word from different speakers often appears in multiple clusters, preventing the model from learning good translations. ZRTools discovers 15,089 clusters in our data, though there are only 10,674 word types. Only 1,614 of the clusters map one-to-one to a unique word type, while a many-to-one mapping of the rest covers only 1,819 gold types (leaving 7,241 gold types with no corresponding cluster).

[5] http://www.nltk.org/

	utterance	call	corpus
Matches	2%	17%	81%
Accuracy	78%	53%	8%

Table 1: UTD matches within utterances, within calls and within the corpus. Matches within an utterance or call are usually from the same speaker.

	utterance split	call split
Oracle	420 (10%)	719 (17%)
Pseudotext	601 (29%)	892 (44%)

Table 2: Number (percent) of out-of-vocabulary (OOV) word tokens or pseudoterms in the test data for different experimental conditions.

Fragmentation of words across clusters renders pseudoterms impossible to translate when they appear only in test and not in training. Table 2 shows that these *pseudotext out-of-vocabulary (OOV)* words are frequent, especially in the call-level split. This reflects differences in acoustic patterns of different speakers, but also in their vocabulary — even the oracle OOV rate is higher in the call-level split.

4.3 UTD is sparse, giving low coverage

UTD is most reliable on long and frequently-repeated patterns, so many spoken words are not represented in the pseudotext, as in Fig. 1. We found that the patterns discovered by ZRTools match only 28% of the audio. This low coverage reduces training data size, affects alignment quality, and adversely affects translation, which is only possible when pseudoterms are present. For almost half the utterances, UTD fails to produce any pseudoterm at all.

5 Speech translation experiments

We evaluate our system by comparing its output to the English translations on the test data. Since it translates only a handful of words in each sentence, BLEU, which measures accuracy of word sequences, is an inappropriate measure of accuracy.[6] Instead we compute precision and recall over

[6] BLEU scores for supervised speech translation systems trained on our data can be found in Kumar et al. (2014).

	source text	gold translation	oracle translation	utd translation
1	cómo anda el plan escolar	how is the school plan going	things whoa mean plan school	school going
2	dile que le mando saludos	tell him that i say hi	tell send best says	say hi
3	sí con dos dientes menos	yeah with two teeth less	two teeth less least	denture yeah teeth
4	o dejando o dejando dos días	or giving or giving two days	improves apart improves apart two days	two days
5	ah ya okey veintitrés de noviembre no	ah yeah okay twenty third of november no	oh ah okay another three fourth november	twenty november

Table 3: Source text (left) paired with translations by humans (gold), oracle, and UTD-based system. Underlined words appear in UTD and the corresponding human translations.

		oracle		pseudotext	
K	metric	utterance	call	utterance	call
1	Prec.	38.6	35.7	7.9	4.0
1	Rec.	33.8	28.4	1.8	0.6
5	Prec.	24.6	23.1	5.9	2.7
5	Rec.	**54.4**	46.4	**5.2**	1.5

Table 4: Precision and recall for $K = 1$ and $K = 5$ under different conditions.

the content words in the translation. We allow the system to guess K words per test pseudoterm, so for each utterance, we compute the number of correct predictions as $corr@K = |pred@K \cap gold|$, where $pred@K$ is the multiset of words predicted using K predictions per pseudoterm and $gold$ is the multiset of content words in the reference translation. For utterances where the reference translation has no content words, we use stop words. The utterance-level scores are then used to compute corpus-level Precision@K and Recall@K.

Table 4 and Fig. 2 show that even the oracle has mediocre precision and recall, indicating the difficulties of training an MT system using only bag-of-content-words on a relatively small corpus. Splitting the data by utterance works somewhat better, since training and test share more vocabulary.

Table 4 and Fig. 2 also show a large gap between the oracle and our system. This is not surprising given the problems with the UTD output discussed in Section 4. In fact, it is encouraging given the small number of discovered terms and the low cluster purity that our system can still correctly translate some words (Table 3). These results are a positive proof of concept, showing that it is possible to discover and translate keywords from audio data even with no ASR or MT system. Nevertheless, UTD quality is clearly a limitation, especially

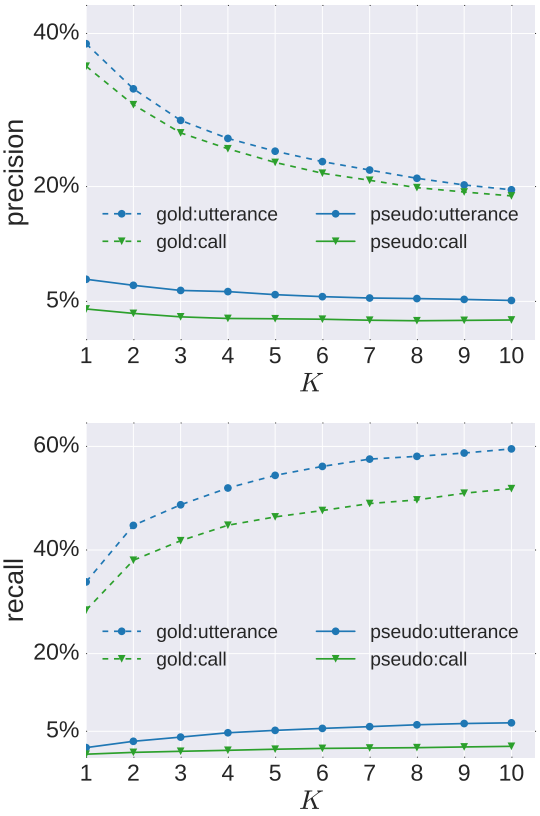

Figure 2: Precision and Recall @K for the call and utterance level test sets.

for the more realistic by-call data split.

6 Conclusions and future work

Our results show that it is possible to build a speech translation system using only source-language audio paired with target-language text, which may be useful in many situations where no other speech technology is available. Our analysis also points to several possible improvements. Poor cross-speaker matches and low audio coverage prevent our system from achieving a high recall, suggesting the of use speech features that are effective in multi-

speaker settings (Kamper et al., 2015; Kamper et al., 2016a) and speaker normalization (Zeghidour et al., 2016). Finally, Bansal et al. (2017) recently showed that UTD can be improved using the translations themselves as a source of information, which suggests joint learning as an attractive area for future work.

On the other hand, poor precision is most likely due to the simplicity of our MT model, and designing a model whose assumptions match our data conditions is an important direction for future work, which may combine our approach with insight from recent, quite different audio-to-translation models (Duong et al., 2016; Anastasopoulos et al., 2016; Adams et al., 2016a; Adams et al., 2016b; Berard et al., 2016). Parameter-sharing using word and acoustic embeddings would allow us to make predictions for OOV pseudoterms by using the nearest in-vocabulary pseudoterm instead.

Acknowledgments

We thank David Chiang and Antonios Anastasopoulos for sharing alignments of the CALLHOME speech and transcripts; Aren Jansen for assistance with ZRTools; and Marco Damonte, Federico Fancellu, Sorcha Gilroy, Ida Szubert, Nikolay Bogoychev, Naomi Saphra, Joana Ribeiro and Clara Vania for comments on previous drafts. This work was supported in part by a James S McDonnell Foundation Scholar Award and a Google faculty research award.

References

Oliver Adams, Graham Neubig, Trevor Cohn, and Steven Bird. 2016a. Learning a translation model from word lattices. In Proc. Interspeech.

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

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

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

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

Laurent Besacier, Bowen Zhou, and Yuqing Gao. 2006. Towards speech translation of non written languages. In Proc. SLT.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. Speech Communication, 56:85–100.

Steven Bird, Ewan Klein, and Edward Loper. 2009. Natural language processing with Python. O'Reilly Media.

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.

Michael A. Carlin, Samuel Thomas, Aren Jansen, and Hynek Hermansky. 2011. Rapid evaluation of speech representations for spoken term discovery. In Proc. Interspeech.

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

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In Proc. ACL.

Aren Jansen and Benjamin Van Durme. 2011. Efficient spoken term discovery using randomized algorithms. In Proc. ASRU.

Herman Kamper, Micha Elsner, Aren Jansen, and Sharon Goldwater. 2015. Unsupervised neural network based feature extraction using weak top-down constraints. In Proc. ICASSP.

Herman Kamper, Aren Jansen, and Sharon Goldwater. 2016a. A segmental framework for fully-unsupervised large-vocabulary speech recognition. arXiv preprint arXiv:1606.06950.

Herman Kamper, Aren Jansen, and Sharon Goldwater. 2016b. Unsupervised word segmentation and lexicon discovery using acoustic word embeddings. IEEE/ACM Trans. Audio, Speech, Language Process., 24(4):669–679.

Gaurav Kumar, Matt Post, Daniel Povey, and Sanjeev Khudanpur. 2014. Some insights from translating conversational telephone speech. In Proc. ICASSP.

Chia-ying Lee, T O'Donnell, and James Glass. 2015. Unsupervised lexicon discovery from acoustic input. Trans. ACL, 3:389–403.

Lara J. Martin, Andrew Wilkinson, Sai Sumanth Miryala, Vivian Robison, and Alan W. Black. 2015. Utterance classification in speech-to-speech translation for zero-resource languages in the hospital administration domain. In Proc. ASRU.

Robert Munro. 2010. Crowdsourced translation for emergency response in Haiti: the global collaboration of local knowledge. In AMTA Workshop on Collaborative Crowdsourcing for Translation.

Alex S. Park and James Glass. 2008. Unsupervised pattern discovery in speech. IEEE Trans. Audio, Speech, Language Process., 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.

Maarten Versteegh, Roland Thiollière, Thomas Schatz, Xuan Nga Cao, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux. 2015. The Zero Resource Speech Challenge 2015. In Proc. Interspeech.

Alex Waibel and Christian Fugen. 2008. Spoken language translation. IEEE Signal Processing Magazine, 3(25):70–79.

Neil Zeghidour, Gabriel Synnaeve, Nicolas Usunier, and Emmanuel Dupoux. 2016. Joint learning of speaker and phonetic similarities with Siamese networks. In Proc. Interspeech.