TOWARDS UNSUPERVISED SPEECH-TO-TEXT TRANSLATION

Yu-An Chung Wei-Hung Weng Schrasing Tong James Glass

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{andyyuan,ckbjimmy,st9,glass}@mit.edu

ABSTRACT

We present a framework for building speech-to-text translation (ST) systems using only monolingual speech and text corpora, in other words, speech utterances from a source language and independent text from a target language. As opposed to traditional cascaded systems and end-to-end architectures, our system does not require any labeled data (i.e., transcribed source audio or parallel source and target text corpora) during training, making it especially applicable to language pairs with very few or even zero bilingual resources. The framework initializes the ST system with a *cross-modal* bilingual dictionary inferred from the monolingual corpora, that maps every source speech segment corresponding to a spoken word to its target text translation. For unseen source speech utterances, the system first performs word-by-word translation on each speech segment in the utterance. The translation is improved by leveraging a language model and a sequence denoising autoencoder to provide prior knowledge about the target language. Experimental results show that our unsupervised system achieves comparable BLEU scores to supervised end-to-end models despite the lack of supervision. We also provide an ablation analysis to examine the utility of each component in our system.

Index Terms— speech-to-text translation, unsupervised speech processing, speech2vec, bilingual lexicon induction

1. INTRODUCTION

Conventional speech-to-text translation (ST) systems, which typically cascade automatic speech recognition (ASR) and machine translation (MT), impose significant requirements on training data. They usually require hundreds of hours of transcribed audio and millions of words of parallel text from the source and target languages to train individual components, which makes it difficult to use this approach on low-resource languages. Although recent works have shown the feasibility of building end-to-end systems that directly translate source speech to target text without using any intermediate source language transcriptions, they still require data in the form of source audio paired with target text translations for end-to-end training [1, 2, 3, 4].

In contrast to ST, which requires paired data for training, recent research in MT has explored fully unsupervised settings—relying only on monolingual corpora from each language. They have shown that unsupervised MT models can achieve comparable (sometimes even superior) results to supervised ones [5, 6]. A key principle behind these unsupervised MT approaches is to initialize a MT model with a bilingual dictionary inferred from monolingual corpora, without using cross-lingual signals [7, 8]. Given a source word, the initial MT model is able to perform word-by-word translation by looking up the dictionary, and can be further improved by leveraging other techniques such as back translation [9].

Recently, [10] showed that these unsupervised bilingual dictionary induction algorithms could also be applied to scenarios where the source and target corpora are of different modalities, namely speech and text. The learned *cross-modal* bilingual dictionary, as we will show in this paper, is capable of performing word-by-word translation, with the difference being that the input, instead of text, is a speech segment corresponding to a spoken word in the source language.

In this paper, we propose a framework for building a ST system using only independent monolingual corpora of speech and text. The two corpora can be collected independently which greatly reduces human labeling efforts. Our framework starts by initializing a ST system with a cross-modal bilingual dictionary inferred from the monolingual corpora to perform word-by-word translation. To further improve the quality of the translations, we incorporate a pre-trained language model (LM) and sequence denoising autoencoder (DAE) [11, 12] that contain prior knowledge about the target language; their primary function is to consider context in lexical choices and handle local reordering and multi-aligned words. To the best of our knowledge, this is the first work that tackles ST in an unsupervised setting. More importantly, experiments show that our unsupervised system achieves comparable results to supervised end-to-end models [3] despite the lack of supervision.

2. PROPOSED FRAMEWORK

Our framework builds on several recently developed techniques for unsupervised speech processing and MT. We first derive a ST system that can perform simple word-by-word translation. Next, we integrate a language model into the framework to introduce contextual information during the translation process. Finally, we post-process the translated results using a DAE to handle local reordering and multi-aligned words. Below we describe each step in detail.

2.1. Word-by-Word Translation System

In our framework, a speech corpus from the source language is first pre-processed using an unsupervised speech segmentation algorithm [13] to generate speech segments corresponding to spoken words. We then apply a neural architecture called Speech2Vec [14, 15] to learn a speech embedding space from the set of speech segments such that each vector corresponds to a word whose semantics has been captured. A text embedding space that captures word semantics can be learned by training Word2Vec [16] on a text corpus from the target language. Based on the assumption that monolingual word embedding spaces are approximately isomorphic, since languages are used to convey thematically similar information in similar contexts [17], it is theoretically possible to align these two spaces.

To achieve this, one can use an unsupervised bilingual dictionary induction (BDI) algorithm to learn a cross-lingual mapping from the source embedding space to the target embedding space. Two of the most representative BDI algorithms are MUSE [7] and VecMap [8], neither of which rely on cross-lingual signals. Note that both these BDI algorithms were originally proposed for aligning two embedding spaces learned from text. In [10], however, the authors showed that MUSE can also be applied to learn a *cross-modal* alignment between embedding spaces learned from speech and text. In our experiments, we include the results of both algorithms for comparison.

We obtain a rudimentary ST system after deriving a cross-modal and cross-lingual mapping from speech to the text corpora, which is essentially a linear transformation W. Given an unseen speech utterance, we first segment it into several speech segments using the speech segmentation algorithm previously mentioned. Then, for each speech segment that potentially corresponds to a spoken word, we map it from the speech embedding space to the text embedding space via W and apply nearest neighbor search to decide its text translation. However, the translations generated by this preliminary system are far from acceptable since nearest neighbor search does not consider the context of the current word. In many cases, the correct translation is not the nearest target word but synonyms or other close words with morphological variations, prompting us to incorporate further improvements.

2.2. Language Model for Context-Aware Beam Search

We incorporate contextual information into word-by-word translation by introducing a LM during the decoding process [18]. Let w_s be the word vector mapped from speech to the text embedding space and w_t the word vector of a possible target word. Given a history h of target words before w_t, the score of w_t being the translation of w_s is computed as:

$$LM(w_t; w_s, h) = \log \frac{f(w_s, w_t) + 1}{2} + \lambda_{\mathrm{LM}} \log p(w_t|h),$$

where λ_{LM} is the weight parameter that decides how *context-aware* the system is, and $f(w_s, w_t) \in [-1, 1]$ is the cosine similarity between w_s and w_t, linearly scaled to the range $[0, 1]$ to make it comparable with the output probability of the LM. Empirically, we found that setting λ_{LM} to 0.1 yields the best performance. Accumulating the scores per position, we perform a beam search to allow only reasonable translation hypotheses.

2.3. Sequence Denoising Autoencoder

We may achieve semantic correctness through learning an appropriate cross-modal bilingual dictionary and using a LM. However, to further improve the quality of the translations, it is also necessary to consider syntactic correctness. To this end, we apply a sequence DAE to correct the translated outputs. By injecting noise to the input sequence during the training process, the DAE learns to output the original (clean) sequence given a corrupted, noisy input. In our framework, we adopt three noise simulation techniques proposed in [18]: word insertion, deletion and permutation. We seek to simulate the noise introduced during the word-by-word translation process with these three techniques. Readers can refer to [18] for more details. Along with the context-aware LM, we found that adopting a DAE further boosts translation performance.

3. EXPERIMENTS

3.1. Datasets

We used an English-to-French speech translation dataset [19] augmented from the LibriSpeech ASR corpus [20]. The dataset is split into train, dev, and test sets; all come with a collection of English speech utterances and their corresponding French text translations. The train set contains 100 hours of speech, which was used to train Speech2Vec [14] to obtain the speech embedding space. For the text embedding space, we trained Word2Vec [16] on two different corpora—the parallel corpus that contains the text translations, and an independent corpus crawled from French Wikipedia. For evaluation, we merged the dev and test sets, resulting in speech data of about 6 hours. BLEU scores [21] were used as the evaluation metric.

3.2. Model Architectures and Training Details

We trained Speech2Vec following the same procedure used in [10]. The text embedding space was trained by Word2Vec using fastText [22] with default settings without subword information. The dimension of both speech and text embeddings is 100. For both VecMap [8] and MUSE [7], we followed the default settings of the implementations released by their original authors. For the LM, we trained a 5-gram count-based LM using KenLM [23] with its default settings. Finally, we implemented the DAE, structured as a 6-layer Transformer [24], with embedding and hidden layer size of 512, a feedforward sublayer size of 2,048, and 8 attention heads.

3.3. Results and Discussions

We first study the similarities between different pairs of embedding spaces to be aligned in Section 3.3.1. We then present the main ST results in Section 3.3.2.

3.3.1. Eigenvector Similarity

Having approximately isomorphic embedding spaces is important for BDI. To quantify whether the embedding spaces are isomorphic, or similar in structure, we computed the eigenvector similarity, which is derived from Laplacian eigenvalues. Both our study and [25] demonstrate that the eigenvector similarity metric is correlated to the performance of the translation task, which implies that the metric reflects the distance between embedding spaces in a meaningful way. The similarity is computed as follows. Let L_1 and L_2 be the Laplacians of two nearest neighbor embedding graphs. We search for the smallest value of k for each graph such that the sum of largest k Laplacian eigenvalues is smaller than 90% of the Laplacian eigenvalues. Then, we select the smallest k across two graphs and compute the squared differences between the largest k Laplacian eigenvalues in two graphs. The differences is the eigenvector similarity we use to measure the similarity between embedding spaces. Note that a *higher* value of the eigenvector similarity metric indicates that the given two embedding spaces are *less* similar.

Table 1 presents the eigenvector similarity of different speech-text pairs. The eigenvector similarity of speech and text embedding space pairs is smaller when we trained the speech embedding using the Speech2Vec algorithm than the Audio2Vec [26] algorithm. These results are expected since Speech2Vec utilizes semantic context of the speech corpus, similarly to how Word2Vec uses that of the text corpus. Furthermore, we applied skip-gram as a training methodology for both algorithms, resulting in isomorphic embedding spaces. In contrast, Audio2Vec focuses on similarities in acoustics rather than semantics, thus the learned embedding space differs fundamentally. Embedding space pairs learned from comparable corpora also yield higher similarity, since the word distributions are more similar; for example, the distribution of English LibriSpeech speech embeddings is more similar to that of the French LibriSpeech text embeddings than French Wikipedia text embeddings.

Table 1: Embedding similarity of different speech and text embeddings pair evaluated by eigenvector similarity. We denote the embedding training method and corpus name in upper and lower case, respectively. For the pair, we denote the speech and text embedding space at the left and right side, respectively. For example, $\mathrm{A}_{\mathrm{libri}}$ - $\mathrm{T}_{\mathrm{wiki}}$ represents the speech embedding space trained on the LibriSpeech corpus using Audio2Vec and the text embedding space trained on Wikipedia corpus. $\mathrm{A}, \mathrm{S}, \mathrm{T}$ indicates Audio2Vec, Speech2Vec and text (Word2Vec) embedding.

Speech & text embedding spaces pair	Eigenvector similarity
$\mathrm{A}_{\mathrm{libri}}$ - $\mathrm{T}_{\mathrm{libri}}$	14.74
$\mathrm{A}_{\mathrm{libri}}$ - $\mathrm{T}_{\mathrm{wiki}}$	15.02
$\mathrm{S}_{\mathrm{libri}}$ - $\mathrm{T}_{\mathrm{libri}}$	6.43
$\mathrm{S}_{\mathrm{libri}}$ - $\mathrm{T}_{\mathrm{wiki}}$	7.17

3.3.2. Speech-to-text Translation

We present the results of our unsupervised approach as well as supervised baselines in Table 2. We trained every system 10 times and report both the best and average performance. In configurations (a-d), we replicate state-of-the-art supervised algorithms and arrived at the conclusion that cascaded systems perform better than their end-to-end counterparts and beam search performs better than greedy search. Note that cascaded systems require more supervision than end-to-end systems, whereas our approach makes no assumptions of having speech-text or language pairs of the comparable corpora.

In configurations (e-l), we showcase the performance of our unsupervised approach, denoted as (BLEU score of VecMap / BLEU score of MUSE) in the columns of Table 2.

Alignment Quality Configurations (e-h) demonstrate that eigenvector similarity of speech and text embedding space pairs have strong positive correlation, namely comparing the relative performances to those shown in Table 1, with the BLEU score of alignment-based ST tasks. The results, from configurations (g) and (h), illustrates that using comparable corpora, and thus better alignment, affects the quality of ST. It also hints that there may exist a threshold of usefulness in alignment performances. Since configurations (e) and (f) lie underneath that threshold, they achieve scores of zero. These findings indicate that eigenvector similarity of embedding spaces could serve as an indicator of unsupervised ST performance.

Unsupervised BDI In all of our unsupervised experiments, we compared the performance between two unsupervised BDI algorithms, VecMap and MUSE. VecMap outper-

Table 2: Different configurations for speech-to-text translation and their performance. The numbers in the section of unsupervised methods denoted as BLEU score (%) of VecMap / BLEU score (%) of MUSE. The notation used in the Table is the same as Table 1. For cascaded systems, we followed the ASR and MT pipeline in [3]. E2E stands for end-to-end.

	System	Best	Average
Cascaded and end-to-end ST systems (supervised)			
(a)	Cascaded + greedy	13.7	13.0
(b)	Cascaded + beam	14.2	13.2
(c)	E2E + greedy	12.3	11.6
(d)	E2E + beam	12.7	12.1
Our alignment-based ST systems (unsupervised)			
(e)	A_{libri} - T_{libri}	0.0 / 0.0	0.0 / 0.0
(f)	A_{libri} - T_{wiki}	0.0 / 0.0	0.0 / 0.0
(g)	S_{libri} - T_{libri}	4.5 / 4.6	4.2 / 2.7
(h)	S_{libri} - T_{wiki}	3.7 / 2.1	3.0 / 0.9
(i)	(g) + LM_{libri}	5.2 / 5.0	4.7 / 2.9
(j)	(g) + LM_{wiki}	9.5 / 8.8	9.0 / 5.7
(k)	(g) + LM_{wiki} + DAE_{wiki}	12.2 / 11.8	11.3 / 7.3
(l)	(h) + LM_{wiki} + DAE_{wiki}	11.5 / 9.1	10.8 / 6.2

forms MUSE in all but one experiment, demonstrating that VecMap can be applied to more difficult scenarios through weak, fully unsupervised initialization with iterative mapping improvements, whereas MUSE, which maps embeddings to the shared space through adversarial training, could only succeed on a more limited set of conditions. Additionally, VecMap trains more stably and faster than MUSE, which has a similar best performance but much lower average performance.

Language Model Integration Integrating a LM improves the performance of ST in all experimental configurations, regardless of the selection of corpus, configurations (g) versus (i) and (j); configurations (h) versus (l) generalize this result to different embedding spaces. By comparing configurations (i) and (j), we discover that the text corpus used to train the LM does not need to be the same as the one used for Word2Vec text embedding space training. In fact, adopting the LM trained on the Wikipedia corpus (LM_{wiki}) produces better performance than using that trained on the LibriSpeech corpus (LM_{libri}). Since introducing the LM grounds words into a context based on the previous word, the much larger LM_{wiki}, containing more words, topic contexts, and sentence structures, serves as a better approximation of the French language than LM_{libri}.

Sequence DAE In configurations (j) versus (k), we show that applying DAE on top of the baseline alignment architecture and LM can further enhance performance in unsupervised ST; the performance is now comparable to end-to-end supervised systems. This also justifies our alignment and post-processing approach since configuration (k) essentially has the same degree of supervision as configurations (c) and (d) and performs similarly well while employing a completely different approach. We attribute this to the DAE's ability to reconstruct corrupted data after translation. Since the semantic alignment method we used may retrieve synonyms based on context, rather than the exact syntactically correct word [10], it is possible that the output even when taking the LM into account is still syntactically incorrect. Moreover, one of the key obstacles in training Speech2Vec lies in the limited performance of unsupervised speech segmentation methods. By incorporating a DAE, we could limit these negative effects after translation. Last but not least, the DAE was trained on LM_{wiki} rather than LM_{libri}. This design decision follows from the observation of the LM corpus choice: since the DAE should learn the French language, a larger, more diverse dataset would perform better than the same dataset used for Word2Vec text embeddings.

Scenario of Real-world ST In configuration (l), we conducted experiments modeling a real-world setting where there exists no comparable speech and text corpora. Instead, we need to collect them independently from different sources. Text data exists in more abundance than speech data and thus we usually adopt the text embedding learned from larger corpus such as Wikipedia, which configuration (h) replicates to our best efforts. By comparing configurations (k) and (l), we demonstrate that the performance of our proposed framework under no supervision is only slightly inferior to the best performance achieved using unsupervised alignment, which requires comparable corpora for speech and text embedding spaces and should be considered supervised. The proposed unsupervised ST framework is thus promising for low language resource ST.

4. CONCLUSIONS

In this paper, we propose a framework capable of performing speech-to-text translation in a completely unsupervised manner. Since the system translates using an inferred cross-modal bilingual dictionary trained without parallel data between speech and text, it could be applied to low or zero-resource languages. By incorporating knowledge of the target language, through adding a LM and a DAE, our system greatly enhances the translation performance: We achieved comparable performance with state-of-the-art end-to-end systems using parallel corpora and only slightly lower scores without it. These results indicate that our approach could serve as a promising first step towards fully unsupervised speech-to-text translation. Future works include testing the proposed framework on other language pairs and examining the relationship between embedding quality and translation performance in more detail.

5. REFERENCES

[1] Ron Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen, "Sequence-to-sequence models can directly translate foreign speech," in *INTERSPEECH*, 2017.

[2] Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater, "Low-resource speech-to-text translation," in *INTERSPEECH*, 2018.

[3] Alexandre Bérard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin, "End-to-end automatic speech translation of audiobooks," in *ICASSP*, 2018.

[4] Alexandre Bérard, Olivier Pietquin, Laurent Besacier, and Christophe Servan, "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*, 2016.

[5] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato, "Phrase-based & neural unsupervised machine translation," in *EMNLP*, 2018.

[6] Mikel Artetxe, Gorka Labaka, and Eneko Agirre, "Unsupervised statistical machine translation," in *EMNLP*, 2018.

[7] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou, "Word translation without parallel data," in *ICLR*, 2018.

[8] Mikel Artetxe, Gorka Labaka, and Eneko Agirre, "A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings," in *ACL*, 2018.

[9] Rico Sennrich, Barry Haddow, and Alexandra Birch, "Improving neural machine translation models with monolingual data," in *ACL*, 2016.

[10] Yu-An Chung, Wei-Hung Weng, Schrasing Tong, and James Glass, "Unsupervised cross-modal alignment of speech and text embedding spaces," in *NIPS*, 2018.

[11] Ilya Sutskever, Oriol Vinyals, and Quoc Le, "Sequence to sequence learning with neural networks," in *NIPS*, 2014.

[12] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol, "Extracting and composing robust features with denoising autoencoders," in *ICML*, 2008.

[13] Herman Kamper, Karen Livescu, and Sharon Goldwater, "An embedded segmental k-means model for unsupervised segmentation and clustering of speech," in *ASRU*, 2017.

[14] Yu-An Chung and James Glass, "Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech," in *INTERSPEECH*, 2018.

[15] Yu-An Chung and James Glass, "Learning word embeddings from speech," in *NIPS Workshop on Machine Learning for Audio Signal Processing*, 2017.

[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS*, 2013.

[17] Antonio Valerio Miceli Barone, "Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders," in *RepL4NLP*, 2016.

[18] Yunsu Kim, Jiahui Gend, and Hermann Ney, "Improving unsupervised word-by-word translation with language model and denoising autoencoder," in *EMNLP*, 2018.

[19] Ali Kocabiyikoglu, Laurent Besacier, and Olivier Kraif, "Augmenting librispeech with french translations: A multimodal corpus for direct speech translation evaluation," in *LREC*, 2018.

[20] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *ICASSP*, 2015.

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: A method for automatic evaluation of machine translation," in *ACL*, 2002.

[22] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[23] Kenneth Heafield, "Kenlm: Faster and smaller language model queries," in *WMT*, 2011.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[25] Anders Søgaard, Sebastian Ruder, and Ivan Vulić, "On the limitations of unsupervised bilingual dictionary induction," in *ACL*, 2018.

[26] Yu-An Chung, Chao-Chung Wu, Chia-Hao Shen, Hung-Yi Lee, and Lin-Shan Lee, "Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder," in *INTERSPEECH*, 2016.