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, Massachusetts, USA

ICASSP
Brighton, UK
May 16, 2019
Outline

• Motivation
• Proposed Framework
• Experiments
• Conclusions
Outline

• Motivation

• Proposed Framework

• Experiments

• Conclusions
Machine Translation (MT)

"the cat is black" → MT system → "le chat est noir"

Automatic Speech Recognition (ASR)

ASR system → "dogs are cute"

Text-to-Speech Synthesis (TTS)

"cats are adorable" → TTS system → (English text, English audio)

Training data pairs

(English text, French translation)

(English audio, English transcription)

(English text, English audio)

Paired data are expensive, but unpaired data are cheap.
Outline

• Motivation

• **Proposed Framework**

• Experiments

• Conclusions
Proposed Framework

• **Goal**: Build a speech-to-text translation system using only unpaired corpora of speech (source) and text (target)

• **Steps at a high-level**
  – Word-by-word translation from source to target language
    * **Unsupervised speech segmentation for segmenting utterances into word segments**
    * **Mapping word segments from speech to text**
  – Improve the word-by-word translation results leveraging prior knowledge on target language
    * **Pre-trained language model**
    * **Pre-trained denoising sequence autoencoder**
Word-by-Word Translation

Testing

“le chat est noir”

French audio corpus

English text corpus

Wikipedia is a multilingual, web-based, free encyclopedia based on a model of openly editable and viewable content, a wiki. It is the largest and most popular …

Do not need to be parallel.

Speech2vec
[Chung & Glass, 2018]

Word2vec
[Mikolov et al., 2013]

Learn a linear mapping $W$ such that

$$W^* = \arg\min_{W \in \mathbb{R}^{d \times d}} ||WX - Y||_F$$

VecMap
[Artexte et al., 2018]
Pre-Trained Language Model

• Word-by-word translation results are not good enough
  – Nearest neighbor search does not consider the context of a word
    * Hubness problem in a high-dimensional embedding space
    * Correct translation can be synonyms or close words with morphological variations

• Language model for context-aware beam search
  – Pre-trained on a target language corpus
  – To take contextual information into account during the decoding process (search)
    * $w_s$: the word vector mapped from the speech to the text embedding space
    * $w_t$: the word vector of a possible target word
    * The score of $w_t$ being the translation of $w_s$ is computed as:

\[
Score(w_s, w_t) = \log \frac{\cos(w_s, w_t)}{2} + 1 + \lambda_{LM} \log p(w_t|h)
\]
Denoising Sequence Autoencoder

• **Goal:** To further improve the translation outcome from the previous step
  – Multi-aligned words
  – Words in wrong orders

• **Denoising autoencoder**
  – Pre-trained on a target language corpus
  – During training, three kinds of artificial noises were added to a clean sentence and the autoencoder was asked output the original clean sentence:
    * **Insertion noise**
    * **Deletion noise**
    * **Reordering noise**

- French sentence #1
  - Word-by-word translation + LM search
  - “Listen me”
  - Denoising

- French sentence #2
  - Word-by-word translation + LM search
  - “Dance me with”
  - Denoising

- French sentence #1
  - Word-by-word translation + LM search
  - “Listen to me”
  - Denoising

- French sentence #2
  - Word-by-word translation + LM search
  - “Dance with me”
  - Denoising
Outline

• Motivation

• Proposed Framework

• Experiments

• Conclusions
Setup

• Data: LibriSpeech English-to-French speech translation dataset¹
  – English utterances (from audiobooks) paired with French translations
    * Speech embedding space: train Speech2vec on the train set speech data (~100 hrs)
    * Text embedding space: train Word2vec on the train set text data vs. crawled French Wikipedia corpus

• Framework components:
  1) Word-by-word translation
     * VecMap² to learn the mapping from speech to text embedding space
  2) Language model for context-aware search
     * KenLM 5-gram count-based LM trained on the crawled French Wikipedia corpus
  3) Denoising sequence autoencoder
     * 6-layer Transformer trained on the crawled French Wikipedia corpus

¹Augmenting LibriSpeech with French translations: A multimodal corpus for direct speech translation evaluation. Kocabiyikoglu et al. 2018
²A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. Artetxe et al. 2018
Setup

• Supervised baselines
  – Cascaded systems
    * Speech recognition + machine translation pipeline (individually trained)
  – End-to-end (E2E) systems
    * A single sequence-to-sequence network w/ attention trained end-to-end

• BLEU scores (%) on the test set (~6 hrs) were reported
  – Both the best and avg. over 10 runs from scratch
Results

<table>
<thead>
<tr>
<th>ST system</th>
<th>Best</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cascaded and end-to-end ST systems (supervised)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a) Cascaded + greedy</td>
<td>13.7</td>
<td>13.0</td>
</tr>
<tr>
<td>(b) Cascaded + beam</td>
<td>14.2</td>
<td>13.2</td>
</tr>
<tr>
<td>(c) E2E + greedy</td>
<td>12.3</td>
<td>11.6</td>
</tr>
<tr>
<td>(d) E2E + beam</td>
<td>12.7</td>
<td>12.1</td>
</tr>
<tr>
<td><strong>Our alignment-based ST systems (unsupervised)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(e) $S_{libri} - T_{libri}$</td>
<td>4.5</td>
<td>4.2</td>
</tr>
<tr>
<td>(f) $S_{libri} - T_{libri} + LM_{wiki}$</td>
<td>9.5</td>
<td>9.0</td>
</tr>
<tr>
<td>(g) $S_{libri} - T_{libri} + LM_{wiki} + DAE_{wiki}$</td>
<td>12.2</td>
<td>11.3</td>
</tr>
<tr>
<td>(h) $S_{libri} - T_{wiki}$</td>
<td>3.7</td>
<td>3.0</td>
</tr>
<tr>
<td>(i) $S_{libri} - T_{wiki} + LM_{wiki} + DAE_{wiki}$</td>
<td>11.5</td>
<td>10.8</td>
</tr>
</tbody>
</table>

Observations:

1. LM and DAE boost translation performance: (e) vs. (f) vs. (g)
2. Domain mismatch affects the alignment quality: (e) vs. (h)
3. Our unsupervised ST is comparable with supervised baselines: (a) ~ (d) vs. (g) and (i)

Unpaired corpora setting
Outline

• Motivation
• Proposed Framework
• Experiments
• Conclusions
Conclusions and Future Work

• An unsupervised speech-to-text framework is proposed
  – Relies only on unpaired speech and text corpora
    * Word-by-word translation
    * Context-aware language model
    * Denoising sequence autoencoder
  – Achieved comparable BLEU scores with supervised baselines
    * Cascaded systems (ASR + MT)
    * End-to-end systems (Seq2seq + attention)

• Improve the alignment quality
• Apply to low-resource languages
• Extend the framework to other sequence transduction tasks (e.g., ASR, TTS)
Thank you!

Questions?