An Attentional Model for Speech Translation Without Transcription

Long Duong¹, Antonios Anastasopoulos², David Chiang², Steven Bird¹, Trevor Cohn¹

¹ University of Melbourne, Australia
² University of Notre Dame, USA
Motivation

Why Speech-based MT?

90% of languages do not have a writing system
Motivation

Why Speech-based MT?

90% of languages do not have a writing system
Motivation

Why Speech-based MT?

90% of languages do not have a writing system

Diagram:

Morgen
Fliege ich
nach Kanada
für Konferenz

Tomorrow
I will fly
to the conference
in Canada
Motivation

Why Speech-based MT?

90% of languages do not have a writing system
Motivation

Low-resource languages

More likely to come with translations rather than transcriptions
Motivation

Endangered languages documentation

Use speech with translations for keyword spotting

Using the Aikuma (Bird 2010) app to collect parallel speech
Speech to Text Translation/Alignment

Our approach

Conventional approach

Automatic speech recognition
+ Speech dictionary
+ Transcribed data

Morgen → fliege → ich → nach Kanada → zur Konferenz

Tomorrow → I will fly → to the conference → in Canada

Machine Translation
+ Large bitext
Speech to Text Translation/Alignment

Our approach

+ Only need Speech-to-text parallel corpus
+ More realistic for low-resource languages

Conventional approach

Automatic speech recognition
+ Speech dictionary
+ Transcribed data

Machine Translation
+ Large bitext

Morgen fliege ich nach Kanada zur Konferenz

Tomorrow I will fly to the conference in Canada
Task Description

**Source side:** Representation of the speech signal

- phone sequence

- frame sequence

**Target side:** Translated text

**Alignment task:** Given source and target side, find best alignment

**Translation task:** Given source side, find best target translation
Creating “Silver” Alignments for Evaluation

CALLHOME Spanish speech and translations

algo de conocimiento

a little bit of knowledge
Creating “Silver” Alignments for Evaluation

CALLHOME Spanish speech and translations

1. Extract features

algo de conocimiento

a little bit of knowledge
Creating “Silver” Alignments for Evaluation

CALLHOME Spanish speech and translations

1. Extract features
2. Force-align speech-transcription

algo de conocimiento

a little bit of knowledge
Creating “Silver” Alignments for Evaluation

CALLHOME Spanish speech and translations

1. Extract features
2. Force-align speech-transcription

algo de conocimiento

a little bit of knowledge
Creating “Silver” Alignments for Evaluation

CALLHOME Spanish speech and translations

1. Extract features

2. Force-align speech-transcription

algo de conocimiento

a little bit of knowledge
Creating “Silver” Alignments for Evaluation

CALLHOME Spanish speech and translations

1. Extract features
2. Force-align speech-transcription
3. Align transcription-translation

algo de conocimiento

a little bit of knowledge
Creating “Silver” Alignments for Evaluation

CALLHOME Spanish speech and translations

1. Extract features
2. Force-align speech-transcription
3. Align transcription-translation
4. Combine the alignments

algo de conocimiento

a little bit of knowledge
Creating “Silver” Alignments for Evaluation

CALLHOME Spanish speech and translations

1. Extract features
2. Force-align speech-transcription
3. Align transcription-translation
4. Combine the alignments

a little bit of knowledge
Creating “Silver” Alignments for Evaluation

CALLHOME Spanish speech and translations

1. Extract features
2. Force-align speech-transcription
3. Align transcription-translation
4. Combine the alignments

~17k utterances
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Wave form
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Wave form
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Wave form
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Wave form
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Wave form
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Waveform
Model Overview

- RNN Decoder
- Attention
- RNN Encoder
- Representation
- Wave form
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Wave form
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Wave form
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Waveform
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Wave form
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Wave form
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Wave form
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Wave form
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Wave form
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Wave form
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Wave form
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Wave form
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Wave form
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Waveform
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Wave form
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Wave form
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Wave form
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Wave form
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Wave form
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Wave form

Tomorrow

<s>
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Waveform
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Wave form
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Wave form
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Wave form

Tomorrow
I
will
fly
Model Overview

RNN Decoder

Attention

RNN Encoder

Representation

Wave form
Model Modifications

Adopted the base Attentional Model (Bahdanau et al. 2014)

+ previous alignment features

+ coverage penalty

+ stacked and pyramidal RNNs

+ alignment smoothing
Previous Alignment Features

Base Attentional Model

\[ \alpha_i = \text{Attend} \left( H_S, H_T^{i-1} \right) \]

Our Attentional Model

\[ \alpha_i = \text{Attend} \left( H_S, H_T^{i-1}, \beta \right) \]

- Position of source RNN
- Position of target RNN
- Last alignment
- Avg of previous alignments
Coverage Penalty

Tomorrow I will fly to Canada

Alignment Matrix

\[ Obj = Obj - \lambda \left( \sum_{j=1}^{m} \sum_{i=1}^{n} \alpha_{ij} - 1 \right)^2 \]
Coverage Penalty

Tomorrow I will fly to Canada

\[ \text{Alignment Matrix} \]

\[
\begin{align*}
\Sigma &= 1 & \Sigma &= 1 & \Sigma &= 1 & \Sigma &= 1 & \Sigma &= 1 & \Sigma &= 1 \\
\Sigma &= 1 & \Sigma &= 1 & \Sigma &= 1 & \Sigma &= 1 & \Sigma &= 1 & \Sigma &= 1 \\
\Sigma &= 1 & \Sigma &= 1 & \Sigma &= 1 & \Sigma &= 1 & \Sigma &= 1 & \Sigma &= 1 \\
\Sigma &= 1 & \Sigma &= 1 & \Sigma &= 1 & \Sigma &= 1 & \Sigma &= 1 & \Sigma &= 1
\end{align*}
\]

\[ \text{Obj} = \text{Obj} - \lambda \left( \sum_{j=1}^{m} \sum_{i=1}^{n} \alpha_{ij} - 1 \right) \]
Stacked and Pyramidal RNN

Encoder: Pyramidal structure (3 layers of LSTM)
- Reduce size by factor of 8
Decoder: 4 LSTM layers
Alignment Smoothing

Motivation: source and target size mismatch

During training smoothing (with $T>1$)

$$
\alpha_{ij} = \frac{\exp(e_{ij}/T)}{\sum_k \exp(e_{ik}/T)}
$$

Tomorrow I will fly to Canada

14
Experiments

Phone-to-Word:
- Represent the speech signal with the phone sequence

Speech-to-Word:
- Represent the speech signal with feature frames
A Simplified Model

RNN Decoder

Attention

RNN Encoder

Representation

Wave form
A Simplified Model

RNN Decoder

Attention

RNN Encoder

Phone Embedding

Phone Sequence

sil a l g o d e k o n o s i m y e n t o sil
Spanish Phones to English Words: Alignment

Alignment performance

- GIZA++: 29.7
- Model 3P (Stahlberg et al, 2012): 31.2
- palign (Neubig et al, 2011): 42.4
- palign (modified): 44.0
- our Model (all mods): 53.6

F-score
Spanish Phones to English Words: Alignment

Feature Ablation

- Basic Attentional Model (Bahdanau et al, 2015) F-score: 42.7
- +alignment features F-score: 46.2
- +coverage penalty F-score: 48.6
- +stacking F-score: 46.3
- +alignment smoothing F-score: 47.3
- +alignment/softmax smoothing F-score: 48.2
- All mods F-score: 53.6
Spanish Phones to English Words: Translation

Translation performance

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>our Model</td>
<td>14.6</td>
</tr>
<tr>
<td>GIZA++</td>
<td>18.2</td>
</tr>
<tr>
<td>pialign (def)</td>
<td>18.9</td>
</tr>
<tr>
<td>pialign (mod)</td>
<td>20.2</td>
</tr>
</tbody>
</table>
Spanish Phones to English Words: Translation

Translation reranking performance

BLEU

GIZA++  |  pialign (def) |  pialign (mod)
--- | --- | ---
18.2 | 18.9 | 20.2
19.9 | 19.8 | 21.1
Keyword : tomorrow

el va mañana para caracas. a qué va a caracas él.
y mañana, y mañana o pasado te voy a poner un paquete.
oh, no, Julio no sé a dónde está y va mañana a caracas, está con richard.
oye, qué bueno, entonces nos vamos tempranito en la mañana.
ono, aquí la gente se acuesta a las dos de la mañana.

Keyword : leave

todo, organizar completo todo, desde los alquileres, la comida, mozo, cantina, todo lo pongo yo aquí.
y entonces dónde lo dejamos pagando estacionamiento y pagando seguro.
sí, el veintiuno. yo salgo de para aquí el dieciséis para florida, y el veintiuno llego a Caracas.
Experiments

Phone-to-Word:
- Represent the speech signal with the phone sequence

Speech-to-Word:
- Represent the speech signal with feature frames
Speech-Word Experiment

RNN Decoder

Attention

RNN Encoder

Representation

Wave form
Speech-Word Experiment

RNN Decoder

Attention

RNN Encoder

PLP features

Wave form
Speech-Word Experiment

Suitable for low-resource scenario.

39 dim. Perceptual Linear Prediction (PLP) features.

Pyramidal RNN structure.

Evaluate using

- Automatic speech recognition.

- Alignment task with simple baseline assuming each English letter aligns to equal number of Spanish frames.
English Speech to English Phones: Transcription

ASR is a sub-problem (without word re-ordering)
- Evaluate on TIMIT dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Phone Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our model (with monotonic constraint)</td>
<td>22.3</td>
</tr>
<tr>
<td>Chorowski et al. (2014)</td>
<td>18.6</td>
</tr>
<tr>
<td>SOTA Graves et al. (2013)</td>
<td>17.7</td>
</tr>
</tbody>
</table>
## Spanish Speech to English Words: Alignment

<table>
<thead>
<tr>
<th>Phones</th>
<th>Aligner</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>Naive baseline</td>
<td>31.7</td>
</tr>
<tr>
<td>gold</td>
<td>Giza++</td>
<td>29.7</td>
</tr>
<tr>
<td>none</td>
<td>AM (all modifications) + Pyramidal RNN</td>
<td>26.4</td>
</tr>
</tbody>
</table>

![Alignment diagram]

**Example alignment:**
- **Input:** I say a little bit of knowledge
- **Output:** Yo digo un poco de conocimiento

**Waveform:**
- **Input:** Yo digo un poco de conocimiento
- **Output:** I say a little bit of knowledge
New task:

Direct modelling of speech+translation data.

Introduced a new dataset for this task.

New model:

Extended the neural attentional model for phones/speech.

Phones to words: +10 alignment F1, +1 translation BLEU