Constructing a Speech Translation System using Simultaneous Interpretation Data

Hiroaki Shimizu, Graham Neubig, Sakriani Sakti, Tomoki Toda, Satoshi Nakamura
Nara Institute of Science and Technology (NAIST), Japan
December 6th 2013
Background

Speech translation

Human interpreters

- What is the matter of inferior?
  - accuracy
  - delay

We focus on the delay problem.
What is the problem of delay?

- Speech translation

When simultaneous interpreters interpret lectures in real time, they perform a **variety of techniques** to shorten the delay.
Techniques of simultaneous interpreters

- **Salami technique** [Jones 02] [Fügen+ 07] [Bangalore+ 12] [Fujita+ 13]
  - Divide longer sentences up into a number of shorter ones

![Diagram](audio-waveform)

- **Adjusting lexical choice**
  - Reduce word reordering

<table>
<thead>
<tr>
<th>English</th>
<th>Japanese</th>
</tr>
</thead>
<tbody>
<tr>
<td>I went to Japan</td>
<td>nihon ni itta</td>
</tr>
<tr>
<td>last year</td>
<td>kyonen</td>
</tr>
</tbody>
</table>

Translator

Simultaneous interpreter

A because B
B dakara A

A because B
A nazenaraba B

2013©Hiroaki Shimizu  AHC-Lab, IS, NAIST
Purpose

Research purpose

Figure out what speech translation can learn from simultaneous interpreters

ST system overall view

Proposed

Simultaneous interpretation data + Translation data

learning

Source sentence → MT system → Target sentence like simultaneous interpreter

Related

[Paulik+ 09]
[Sridhar+ 13]
Overview

1) Collecting simultaneous interpretation data

Simultaneous interpretation data

Source sentence

MT system

Target sentence like simultaneous interpreter

2) Difference between simultaneous interpretation and translation data

Translation data

Learning

3) Using the simultaneous interpretation data

4) Experiment and Result
Simultaneous interpretation data

- **Materials**
  - TED (English → Japanese)

  Possible to compare translated subtitles with simultaneous interpretation data

- **Interpreters**
  - Three simultaneous interpreters
  - Different experience levels

<table>
<thead>
<tr>
<th>Experience</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 years</td>
<td>S rank</td>
</tr>
<tr>
<td>4 years</td>
<td>A rank</td>
</tr>
<tr>
<td>1 year</td>
<td>B rank</td>
</tr>
</tbody>
</table>

  Allow us to compare characteristics of interpreters of different levels
Overview

1) Collecting simultaneous interpretation data

2) Difference between simultaneous interpretation and translation data

3) Using the simultaneous interpretation data

4) Experiment and Result
Difference between translation data and simultaneous interpretation data

Motivation

We compare translation data with the simultaneous interpretation data to find the difference.
We hypothesize the similarities of T1-T2 and I1-I2 are higher than any other combinations.
Result: difference simultaneous interpretation data and translation data

<table>
<thead>
<tr>
<th>BLEU</th>
<th>Translator - TED</th>
<th>TED - S rank</th>
<th>Translator - S rank</th>
<th>S rank - A rank</th>
<th>Translator - A rank</th>
<th>TED - A rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>19.18</td>
<td>71.39</td>
<td>61.6</td>
<td>59.7</td>
<td>52.51</td>
<td>49.36</td>
<td>49.4</td>
</tr>
<tr>
<td>13.17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.62</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Translation data pairs are highest in all combinations.
- Translation and simultaneous interpretation data pairs are lower than translation data pair.
Simultaneous interpretation data pair is unexpectedly low.
The reason that simultaneous interpretation data pair is unexpectedly low

<table>
<thead>
<tr>
<th>Data</th>
<th>Words (Ja)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation</td>
<td></td>
</tr>
<tr>
<td>Translator</td>
<td>4.58k</td>
</tr>
<tr>
<td>TED subtitle</td>
<td>4.64k</td>
</tr>
<tr>
<td>Simultaneous interpretation</td>
<td></td>
</tr>
<tr>
<td>S rank</td>
<td>4.44k</td>
</tr>
<tr>
<td>A rank</td>
<td>3.67k</td>
</tr>
</tbody>
</table>

- A rank is more similar to S rank than any others

Translation data and simultaneous interpretation data are different from the view of the similarity measures.
Overview

1) Collecting simultaneous interpretation data

2) Difference between simultaneous interpretation and translation data

Simultaneous interpretation data + Translation data → Learning

Source sentence → MT system → Target sentence like simultaneous interpreter

3) Using the simultaneous interpretation data

4) Experiment and Result
Learning of the MT system

We use simultaneous interpretation data for three steps

- **Tuning (Tu)**
  - the parameters such as the reordering probabilities and word penalty to learn the style of simultaneous interpreters.

- **Language model (LM): linear interpolation**
  - The word order and lexical choice of translation is similar to simultaneous interpretation.

- **Translation model (TM): fill-up [Bisazza+ 11]**
  - Like LM, lexical choice is similar to simultaneous interpretation.
Overview

1) Collecting simultaneous interpretation data

2) Difference between simultaneous interpretation and translation data

3) Using the simultaneous interpretation data

4) Experiment and result
## Data

### Task

- TED talks (English → Japanese)

<table>
<thead>
<tr>
<th></th>
<th>Translation data</th>
<th>Simultaneous interpretation data</th>
</tr>
</thead>
<tbody>
<tr>
<td>TM, LM (en/ja)</td>
<td>1.57M / 2.24M</td>
<td>29.7k / 33.9k</td>
</tr>
<tr>
<td>Tune (en/ja)</td>
<td>12.9k / 19.1k</td>
<td>12.9k / 16.1k</td>
</tr>
<tr>
<td>Test (en/ja)</td>
<td>-</td>
<td>11.5k / 14.9k</td>
</tr>
</tbody>
</table>

1) Using only the data from the S rank interpreter

2) Simultaneous interpretation data is used for reference NOT translation data
Setup

- **Automatic sentence segmentation method**
  - Dividing method using right probability [Fujita+ 13]

- **Evaluation method**
  1) Translation accuracy
     - BLEU, RIBES
  2) Delay
     - Time from start of input to completion of translation

(100% accurate ASR and do not consider speech synthesis)
Result: learning of the MT system (BLEU)

- Better performance
- Phrase unit
- Sentence unit
- Delay (sec)

Similar to simultaneous interpreter

Shorten the delay
Result: learning of the MT system (BLEU)

More similar to simultaneous interpreters
Proposed system does not show improvement from the view for RIBES, because tuning is optimized for BLEU.
## Example of translation results

<table>
<thead>
<tr>
<th></th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Src</strong></td>
<td>If you look at in the context of the history you can see what this is doing</td>
</tr>
<tr>
<td><strong>Ref</strong></td>
<td>過去から流れを見てみますと/災害は/このように/増えています from the past / look at the context and / disasters are / like this increasing</td>
</tr>
<tr>
<td><strong>Baseline</strong></td>
<td>見てみると/歴史の中で/見ることができます/これがやっていること looking at / in the history / you can see / what this is doing</td>
</tr>
<tr>
<td><strong>Proposed</strong></td>
<td>では/歴史の中で/見ることができます/これがやっていること ok / in the history / you can see / what this is doing</td>
</tr>
</tbody>
</table>

- Choose shorter phrase to reduce the number of the words
- Start a sentence with the word “and” (over 25% sentence)
Setup: comparing the system with human simultaneous interpreters

We compare our proposed system with the human simultaneous interpreters

- Compare with the human simultaneous interpreters
  - A rank (4 year)
  - B rank (1 year)

- We use ASR results as input to the translation system
  - WER is 19.36%
Result: comparing the system with human simultaneous interpreters (BLEU)

The system achieves result slightly lower than human simultaneous interpreters from the view of BLEU.
Result: comparing the system with human simultaneous interpreters (RIBES)

From the view of RIBES, the system and B rank (1 year) interpreter achieve similar result.
Conclusion

● Purpose
  – Generate translations similar to those of a simultaneous interpreter

● Proposed
  – Use simultaneous interpretation data for learning

● Result
  – Output is more similar to simultaneous interpreter

● Future works
  – Subjective evaluation
Thank you!
Questions?
Appendix
Question list
The reason that simultaneous interpretation data pair is unexpectedly low

<table>
<thead>
<tr>
<th>Data</th>
<th>Words (Ja)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation</td>
<td></td>
</tr>
<tr>
<td>Translator</td>
<td>4.58k</td>
</tr>
<tr>
<td>TED subtitle</td>
<td>4.64k</td>
</tr>
<tr>
<td>Simultaneous</td>
<td></td>
</tr>
<tr>
<td>interpretation</td>
<td></td>
</tr>
<tr>
<td>S rank</td>
<td>4.44k</td>
</tr>
<tr>
<td>A rank</td>
<td>3.67k</td>
</tr>
</tbody>
</table>

S rank can interpret, but A rank cannot.

A rank is more similar to S rank than any others

Translation data and simultaneous interpretation data are different from the view of the similarity measures
There is no difference to use the simultaneous interpretation data for learning right probability.
Result: learning of translation timing (RIBES)

There is no difference to use the simultaneous interpretation data for learning right probability.
Why English-Japanese Difficult?

In 25 years it is gone from this to this

25 年で このような形 から このような形 に なりました

More difficult to divide the sentence with keeping the accuracy at English-Japanese
Evaluation method

Delay

\[ D = U + T \]

- \( U \): Waiting time before we can start translating
- \( T \): Time required for MT decoding
Right probability [Fujita+ 13]
Why BLEU is quite low?