VISA: An Ambiguous Subtitles Dataset for Visual Scene-Aware Machine Translation

Yihang Li, Shuichiro Shimizu, Weiqi Gu, Chenhui Chu, Sadao Kurohashi

1. Introduction

- Lack of language ambiguity in existing multimodal machine translation (MMT) datasets [Caglayan et al., 2019] -- Do not require visual information
- We construct a new dataset VISA -- 40k Japanese--English parallel subtitles -- Corresponding video clips
- The dataset has following key features:
  1. Subtitles from movies and TV episodes
  2. Ambiguous source subtitles
  3. Divided into Polysemy and Omission
- We conduct experiments on the VISA dataset with the latest video-guided machine translation (VMT) architecture to set a baseline for the dataset

2. Pipeline

3. Data Example

4. Splits

<table>
<thead>
<tr>
<th>Split</th>
<th>Train</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polysemy</td>
<td>18,666</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>Omission</td>
<td>17,214</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>Combined</td>
<td>35,880</td>
<td>2,000</td>
<td>2,000</td>
</tr>
</tbody>
</table>

5. Experiment Settings

- Datasets -- Polysemy part, Omission part, the whole VISA dataset
- Models -- 4 models based on the VMT architecture described in Gu et al. (2021)
- Metrics -- BLEU, METEOR, RIBES

6. Results and Discussion

- VMT works better on Omission while NMT works better on others
- Why doesn’t the current VMT model work well on VISA?
  - The videos do not necessarily contribute to the disambiguation
  - Lack of speaker recognition
  - Model can not capture emotional information