Hand in Hand: Automatic Sign Language to English Translation
Daniel Stein, Philippe Dreuw, Hermann Ney
Human Language Technology and Pattern Recognition, RWTH Aachen University
Sara Morrissey, Andy Way
National Centre for Language Technology, Dublin City University

Introduction

- problem: all approaches in sign language research work on an intermediate language
- sign language in machine recognition:
  - input: video of signing person
  - output: semantic sign language representation (e.g. glosses)
- sign language in machine translation:
  - input: semantic sign language representation
  - output: written language (i.e. English)
- not directly intelligible by either hearing or deaf people
- incorporating statistical machine translation (SMT) on top of the recognition process:
  - converts glosses into written English
  - works even for very small corpora
- data derived during the recognition can be used as additional knowledge source

Intermediate Notation

- sign languages lack a formally adopted writing system
- syntactic representations describe handshape, location and movement of a sign
- glosses are a semantic representation of sign language
- conventionally transcribed in the upper case stem form of the local spoken language
- includes spatial and non-manual information

Sign Language Translation

- state-of-the-art phrase-based statistical machine translation system
  - for a recognized sequence $f'_j$ we maximize a translation probability for target sentences $c_t$
  - log-linear combination model:
    $$ p(c_t | f'_j) = \exp \left( \sum_{m=1}^{M} \lambda_m h_m(c_t, f'_j) \right) / \sum_{c'_t} \exp \left( \sum_{m=1}^{M} \lambda_m h_m(c'_t, f'_j) \right) $$
  - set of different features $h_m$, scaling factors $\lambda_m$
  - trained with downhill simplex algorithm
  - tracking positions of the sentences were clustered and their mean calculated
  - for deictic signs, the nearest cluster according to the Euclidean distance was added as additional word information for the translation model

RWTH-Boston-104 database:

- 161 training sentences, 40 test sentences

<table>
<thead>
<tr>
<th>Experimental Results</th>
<th>WER[%]</th>
<th>PER[%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>recognition</td>
<td>17.9</td>
<td>-</td>
</tr>
<tr>
<td>translation</td>
<td>21.2</td>
<td>20.1</td>
</tr>
<tr>
<td>sign-to-speech</td>
<td>27.6</td>
<td>25.6</td>
</tr>
</tbody>
</table>

RWTH-Boston-Hand database:

- 1000 annotated frames, 2.3% tracking error rate
- tracking of head and dominant-hand for ASLR

<table>
<thead>
<tr>
<th>Translation Features (different split)</th>
<th>WER[%]</th>
<th>PER[%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>without tracking</td>
<td>28.9</td>
<td>23.8</td>
</tr>
<tr>
<td>with tracking</td>
<td>26.5</td>
<td>23.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Translation Example</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>without tracking</td>
<td>John gives that man a coat.</td>
<td></td>
</tr>
<tr>
<td>with tracking</td>
<td>John gives the man over there a coat.</td>
<td></td>
</tr>
</tbody>
</table>

Conclusion

- first data-driven automatic sign-language-to-speech translation system
- approach works for extremely small corpora typically encountered
- can be easily trained on new language pairs and new domains
- incorporation of the tracking data for the deictic words helps the translation system to discriminate between
  - distinctive article,
  - locative reference or
  - discourse entity reference

Outlook

- stemming of the glosses (i.e. leaving out the inflection)
- adding relevant features later in the translation
- model for all discourse entities
- handling spatial verb flexion, time information