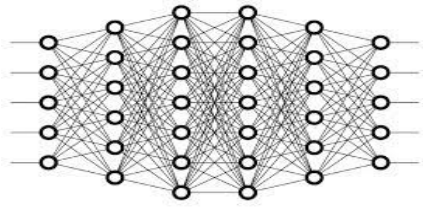
Sign Language Production using Deep Learning

author: Alua Musralina
advisor: Radoslav Neychev

Phystech School of Applied Mathematics and Informatics
Moscow Institute of Physics and Technology, Russia

musralina.az@phystech.edu
www.mipt.ru

Moscow, Russia
May 28, 2023



Outline

1. Motivation
2. Tasks description
3. Dataset for Sign Language Production
4. Methods overview
5. Gloss Generation results
6. Keypoints generation
7. Conclusion
8. Outlook

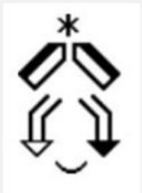
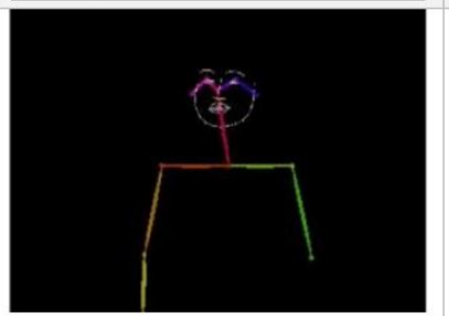
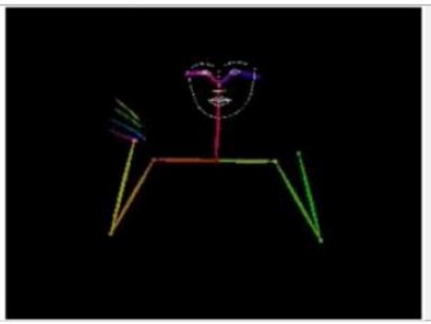
Motivation

The Sign language is a language, that is visually performed with hand gestures, body postures and face expressions, so the meaning of signs depends on the combination of all of them.

1. There are up-to 300 different signed languages (United Nations 2022) and up-to 70 million deaf people exist in our world (World Federation of the Deaf 2022).
2. There are two main tasks can be implemented for Translation:
   - 2.1 from sign language (Sign Language Recognition)
   - 2.2 into sign language **(Sign Language Production)**

Motivation

Every sign has an unique identifier, which is called as **Gloss**. There is no direct alignment between sign sequences and spoken language sentences.

English Translation	Gloss	Notation	Pose	Video/Avatar
House	HOUSE			
What's the matter? What's wrong?	Wrong-What			
Different But	DIFFERENT BUT			

https://research.sign.mt

Tasks

Tasks:

- Methods overview
  - end-to-end system (text2sign architecture)
  - with the intermediate results production (text2gloss architecture)
- Implementation and evaluation of Gloss production
  - seq2seq from scratch
  - seq2seq with pre-trained model
- Implementation of mapping from Gloss to Sign Keypoints

Goal:

Tasks

Tasks:

- Methods overview
  - end-to-end system (text2sign architecture)
  - with the intermediate results production (text2gloss architecture)
- Implementation and evaluation of Gloss production
  - seq2seq from scratch
  - seq2seq with pre-trained model
- Implementation of mapping from Gloss to Sign Keypoints

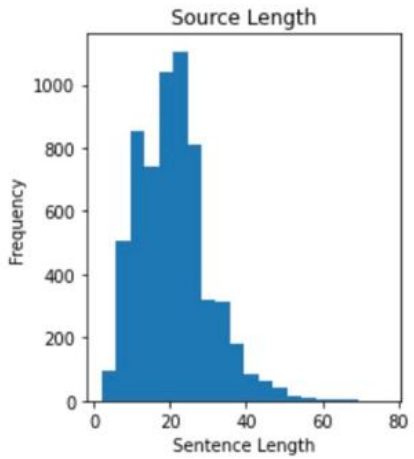
Goal:

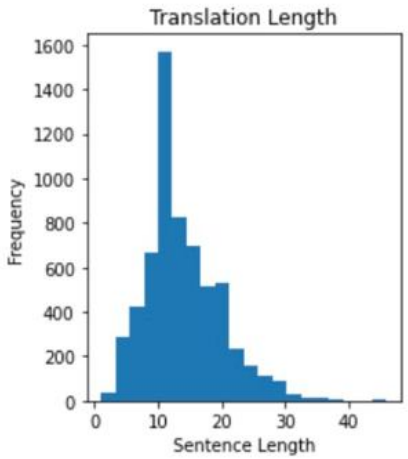
- Implementation and evaluation of possible improvements of seq2seq models for text2gloss translation for Sign Language Production.

Phoenix Dataset for SLP

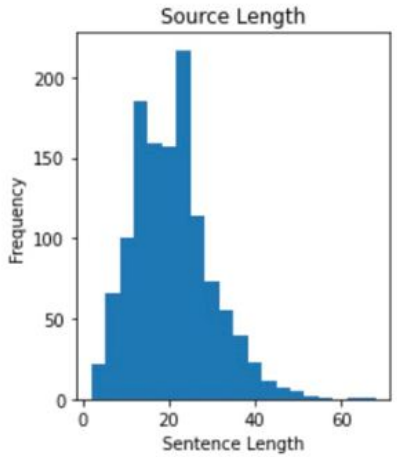
	src	trg
0	tiefer luftdruck bestimmt in den nächsten tage...	DRUCK TIEF KOMMEN
1	das bedeutet viele wolken und immer wieder zum...	ES-BEDEUTET VIEL WOLKE UND KOENNEN REGEN GEWIT...
2	meist weht nur ein schwacher wind aus untersch...	WIND MAESSIG SCHWACH REGION WENN GEWITTER WIND...

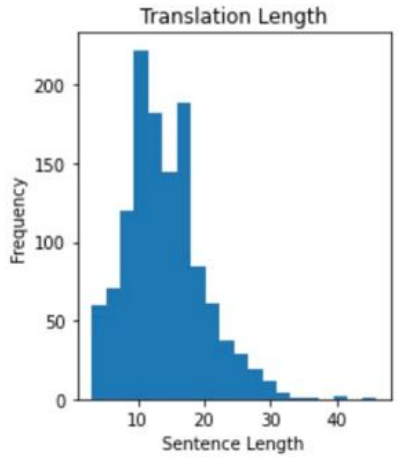
```
Length distribution in Train data
Average length of source sentences: 20.69
Average length of target sentences: 13.71
```






```
Length distribution in Test data
Average length of source sentences: 20.81
Average length of target sentences: 14.02
```





8257 videos of 9 different signers are provided, with a vocabulary of 2887 German words and 1066 different sign glosses.

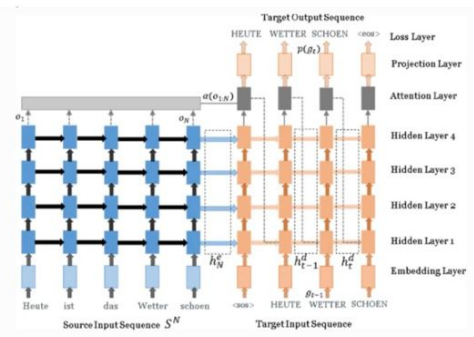
Methods overview

In general, text to sign language process can be divided into three modules:

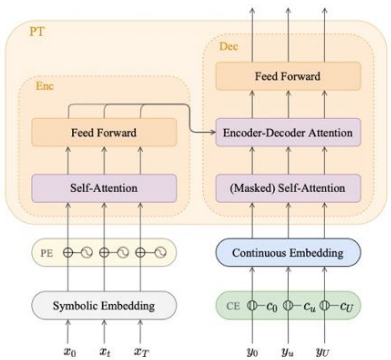
- pre-processing the input text
- convert into sign sequences using MT
- generate videos or animated avatars

Methods examples overview:

- Production of concatenated isolated signs (Stephanie Stoll).
- Automatically Mapping between the text and the pose sequences, without need of Gloss production. Glosses are produced only for comparison (Ben Saunders).



NMT-based encoder-decoder architecture with Luong attention (Stoll)



Progressive Transformer (Saunders)

Evaluation of Gloss production

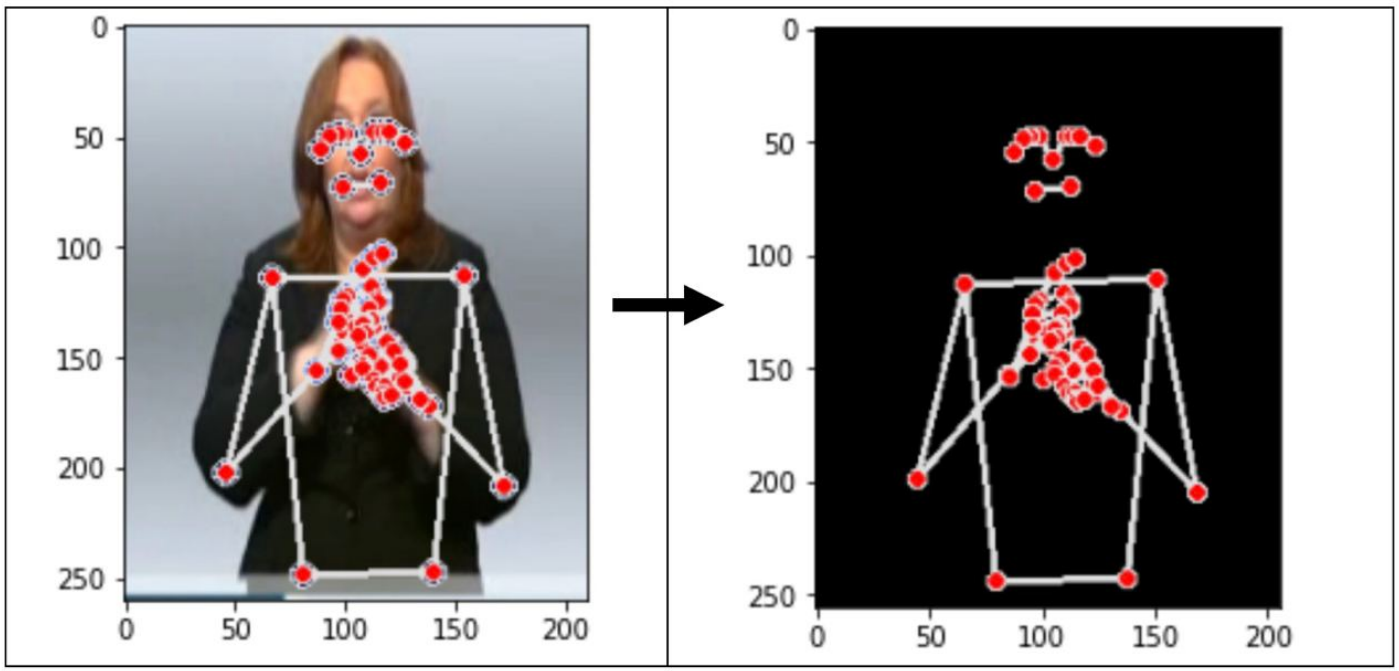
Text to Gloss production results:

- with small dataset (ca. 8000 pairs) and relatively simple model a good BLEU 23.72 for generating the Gloss is achieved, which further could be used for avatar generation.
  Methods results comparison table:

Method	Number of parameters	Bleu-4	Rouge-1
Stoll: Enc-Dec with attention	Not available	19.10	54.55
Saunder: Progressive transformer case, text2gloss part	Not available	15.26	48.10
Our: Enc-Dec without attention and with LSTM	2,452,438	7.35	29.15
Our: Encoder-Decoder model with attention and with bidirectional GRU	4,937,430	19	49.69
Our: pre-trained Huggingface transformer	73, 886, 208	23.72	64.92

Keypoints generation

- 2D upper body joint and facial landmark positions are first extracted using MediaPipe library:
  - 21 keypoints for each hand
  - 33 keypoints for pose
- Lookup table is created for each gloss.
- The glosses are mapped to the sequence of prepared frames (video).



Outlook

- Further fine-tune the Huggingface transformer
- Generate more natural movements of avatars (out of the scope of this work)