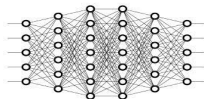# Sign Language Production using Deep Learning

author: Alua Musralina
supervisor: Radoslav Neychev

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

musralina.az@phystech.edu
www.mipt.ru
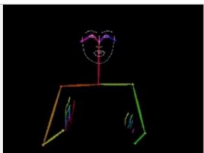
Moscow, Russia
June 24, 2023
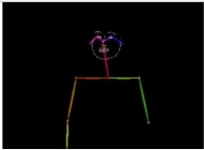
Outline

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)**

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:

- ▶ Methods overview
  - end-to-end system
  - sign generation with the intermediate results production
- ▶ Implementation and evaluation of Gloss production:
  - using seq2seq models from scratch
  - fine-tuning a pre-trained model
- ▶ Implementation of mapping from Gloss to Sign

Goal:

Tasks:

- ▶ Methods overview
    - end-to-end system
    - sign generation with the intermediate results production
- ▶ Implementation and evaluation of Gloss production:
    - using seq2seq models from scratch
    - fine-tuning a pre-trained model
- ▶ Implementation of mapping from Gloss to Sign

Goal:

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

German sentences (Source)	sequence of Glosses (Target)
tiefer luftdruck bestimmt in den nächsten tagen unser wetter	DRUCK TIEF KOMMEN
das bedeutet viele wolken und immer wieder zum teil kräftige schauer und gewitter	ES-BEDEUTET VIEL WOLKE UND KOENNEN REGEN GEWITTER KOENNEN
meist weht nur ein schwacher wind aus unterschiedlichen richtungen der bei schauern und gewittern stark böig sein kann	WIND MAESSIG SCHWACH REGION WENN GEWITTER WIND KOENNEN
am mittwoch hier und da nieselregen in der nordwesthälfte an den küsten kräftiger wind	MITTWOCH REGEN KOENNEN NORDWEST WAHRSCHEINLICH NORD STARK WIND
und nun die wettervorhersage für morgen freitag den sechsten mai	JETZT WETTER WIE-AUSSEHEN MORGEN FREITAG SECHSTE MAI ZEIGEN-BILDSCHIRM

Average Source Sentence Length: 17.45649541601799
Average Target Sentence Length: 13.75748140460128
Length distribution in Train data
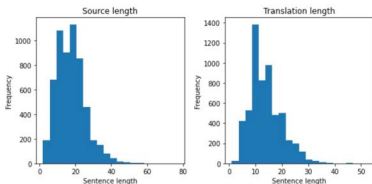
Average Source Sentence Length: 17.67043618739903
Average Target Sentence Length: 13.962843295638127
Length distribution in Test data
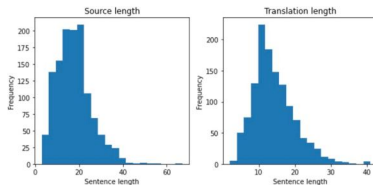
8257 videos of 9 different signers are provided, with a vocabulary of 2887 German words and 1066 different sign glosses.

Sign Language Production steps

In general, Sign Language Production can be divided into three modules:

- ▶ pre-processing the input text (tokenization)
- ▶ convert into Sign or Gloss sequences using NMT
    - seq2seq architecture for text2gloss production
    - Progressive Transformer for direct text2sign production
- ▶ generate videos or animated avatars
    - The Hamburg Notation System for Sign Languages (HamNoSys) - formally describes the position of the hands, facial expressions and body in space for sign languages.
    - mapping from Gloss into Sign using the dataset for production.

Text2Gloss vs. Text2Gloss2Sign

- ▶ 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).

Architecture	BLEU-4	ROUGE-L
T2P	10.51	32.46
T2G2P	9.68	31.07
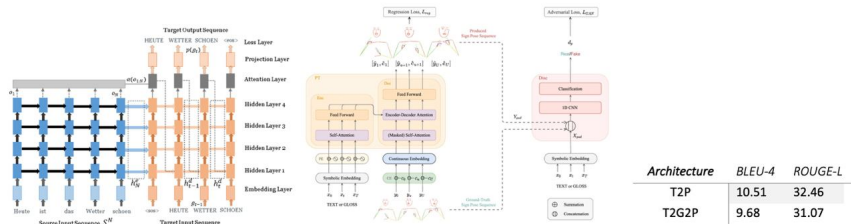
NMT-based encoder-decoder architecture with Luong attention (Stoll)	**Progressive Transformer (Saunders)**

Gloss generation evaluation results

Metrics:

- **BLEU** (Bilingual Evaluation Understudy) "the closer a machine translation is to a professional human translation, the better it is".
- **ROUGE-L** (Recall Oriented Understudy for Gisting Evaluation) - uses LCS (Longest Common Subsequence), where precision and recall values are calculated based on the length of the longest common subsequence between the candidate and reference summaries.

Methods results comparison table:

Method	Number of parameters	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L F1
Stoll: Enc-Dec with attention	Not available	50.67	32.25	21.54	15.26	48.10
Saunder: Progressive transformer, text2gloss part	Not available	55.18	37.10	26.24	19.10	54.55
Our: Enc-Dec model without attention and with LSTM	2,436,911	26.90	16.80	11.50	8.27	31.42
Our: Enc-Dec model with attention and with bidirectional GRU	3,053,743	39.27	28.54	21.79	17.04	44.81
Our: fine-tuned pre-trained Huggingface transformer	73,886,208	64.14	45.11	32.47	24.27	61.35

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 keypoints from table

Demo: Text2Sign system input text example: "das bedeutet viele Wolken und immer wieder zum teil kräftige Schauer und Gewitter". Output will be presented in Demo-Video.
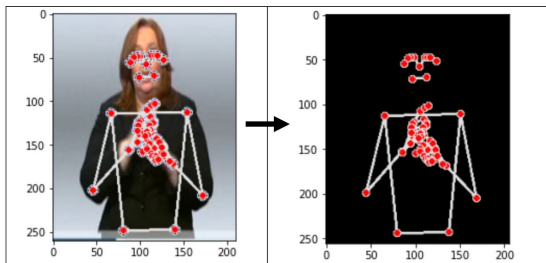
Conclusion and outlook

Conclusion:

- ▶ Pre-trained Transformer dedicated for NMT for German-English languages translations, could be fine-tuned and applied for German-Sign languages translations.
- ▶ With small dataset (ca. 8000 pairs) and relatively simple model a good BLEU 24.27 for generating the Glosses is achieved, which further could be used for avatar generation.
- ▶ Keypoints generation from the Glosses depends on the parallel corpus dedicated for the SLP. Parallel corpus can be prepared using the recognition task.

Outlook:

- ▶ Prepare the corpus for SLP as a results of solving the recognition task.
- ▶ Solve the challenges as not smooth and not natural movements.