Changing the Representation: Examining Language Representation for Neural Sign Language Production

SLTAT 2022

Harry Walsh, Ben Saunders, Richard Bowden

Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey

1. Introduction

Sign Language Production (SLP) traditionally consists of two steps:
1. Translating from a spoken language sentence to a sequence of glosses.
2. Producing a sign language video given a sequence of glosses.

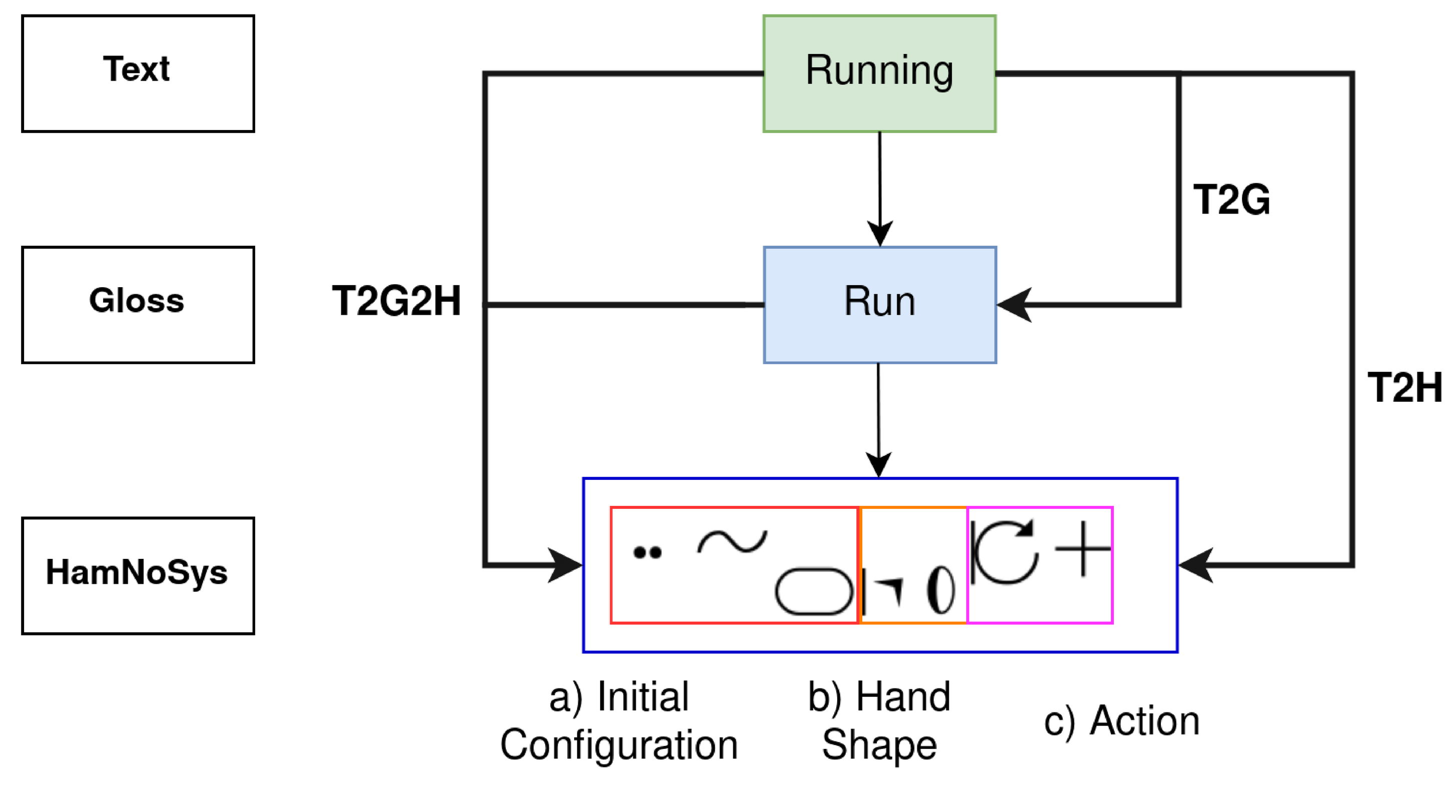
In this paper we apply Natural Language Processing (NLP) techniques to the first step of the SLP pipeline.

Contributions:
We find the following,

- BPE is the best tokenizer.
- Hand shape can be used as additional supervision during training.
- BERT can be used to create better sentence level embedding for Text to HamNoSys (T2H).

2. Representation

Given a **Spoken Language Sentence**, $\mathcal{X} = (x_1, ..., x_W)$ with W words, our model aims to produce a sequence of **glosses**, $y = (y_1, y_2, ..., y_G)$ with G glosses (T2G), or a sequence of **HamNoSys**, $z = (z_1, z_2, ..., z_H)$ with H symbols (T2H).

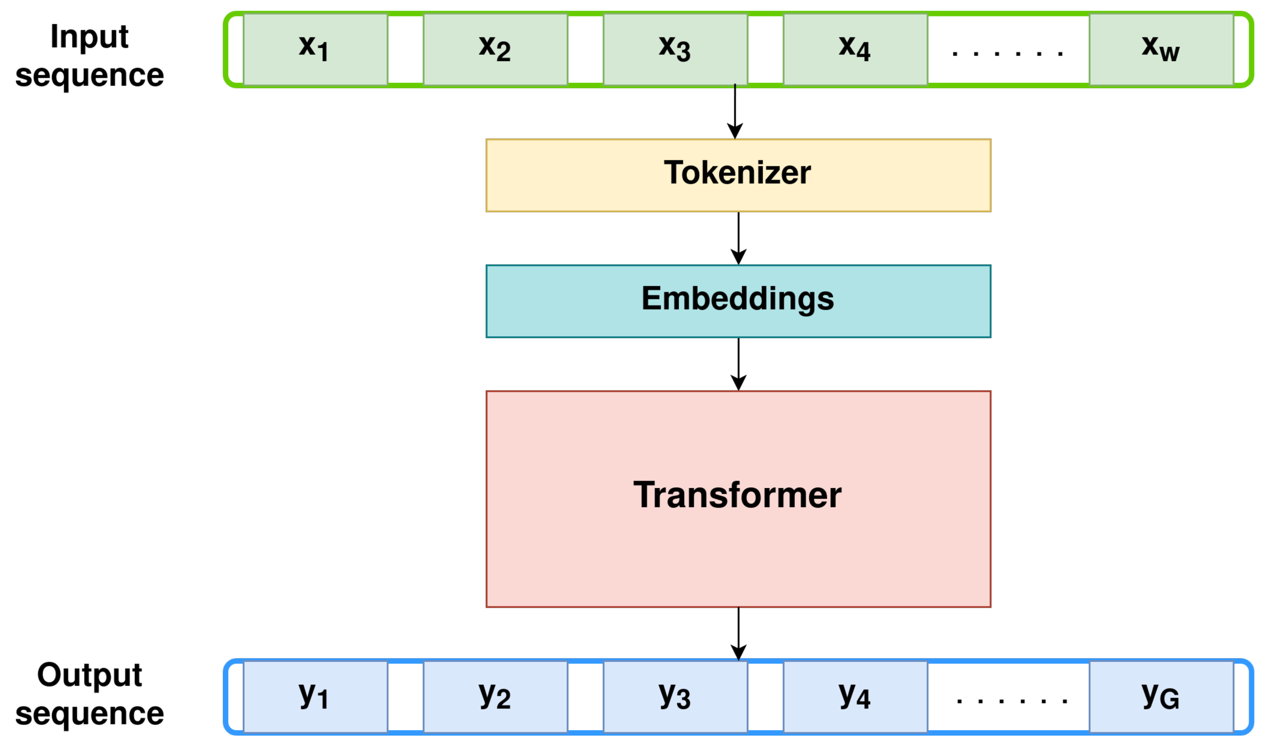


What is Gloss? Gloss is the written word associated with a sign.

What is HamNoSys? HamNoSys can be considered to be a phonetic representation of Sign Language. Each symbol of HamNoSys describes a different component of a Sign. Each sign of HamNoSys has three core components; initial configuration, hand shape and action.

3. General Model Architecture

Using a traditional encoder-decoder transformer and we experiment with changing the output sequence **representation** (2.), **tokenizer** (4A.) and **embedding** methods (4B.). Furthermore we add HamNoSys hand shape as an additional **supervision** during training (4C.).

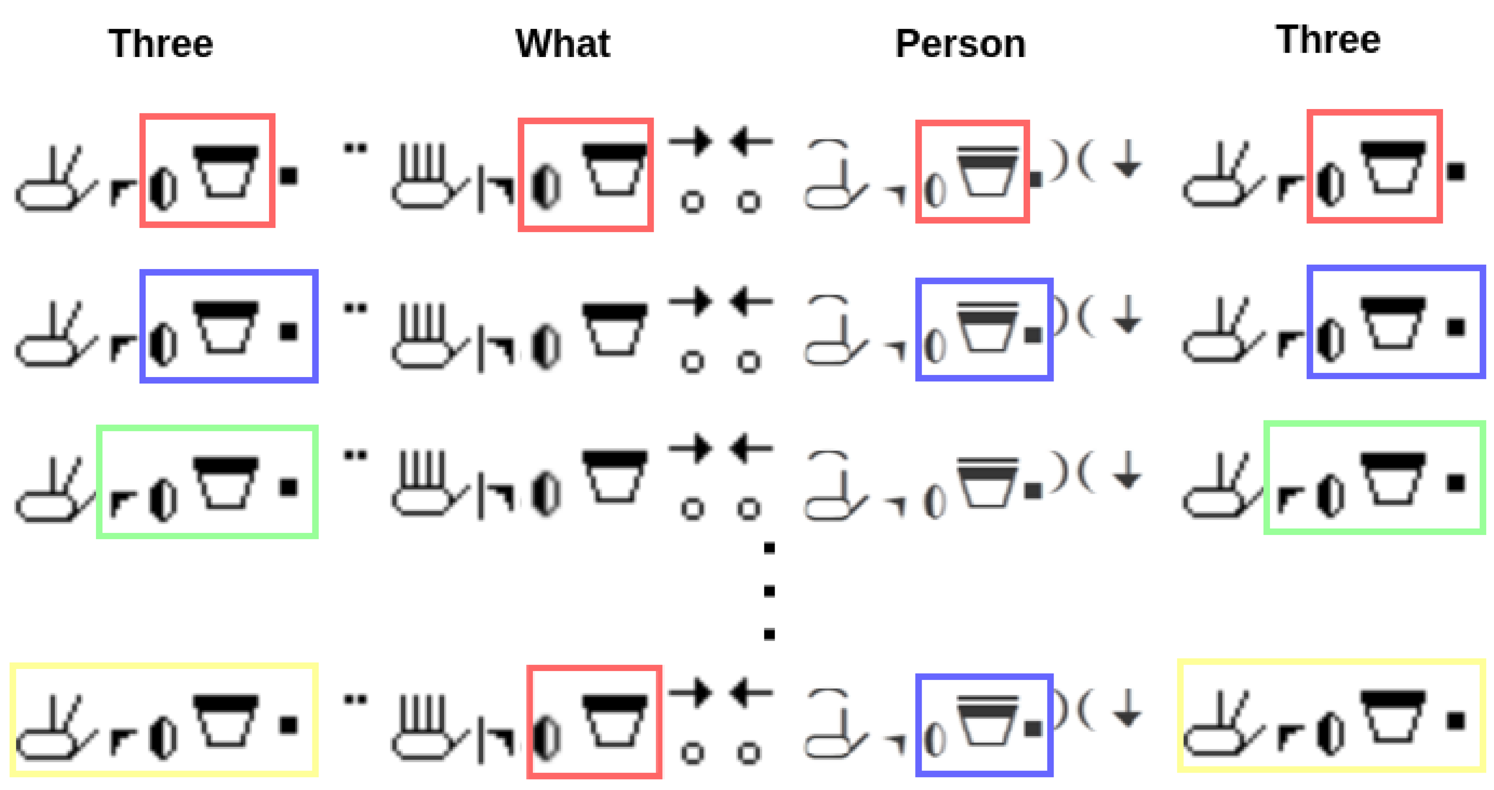


4A. Tokenizers

Tokenization is the process of breaking the input sequence into smaller chunks.
We use several tokenizers;

- **Word** - segmented by white space.
- **Character** - segmented by each character.
- **BPE** - segmented by the most commonly occurring sequential characters.

An example of BPE being applied to HamNoSys:



Best tokenizer combination:
T2G - Input: Word, Output: BPE - **22.06** BLEU-4
T2H - Input: BPE, Output: BPE - **26.14** BLEU-4

4B. Embeddings

The input sequence x is first tokenized then embedded by projecting the sequence into a continuous space. We experiment with three different embedding techniques;

- Linear layer
- BERT
- Word2Vec

Best input embedding:
T2G - Linear Layer - **16.24** BLEU-4
T2H - BERT - **20.26** BLEU-4

4C. Additional Supervision

There exists a strong correlation between hand shape and meaning. We investigate forcing the transformer to predict the hand shape alongside the gloss or HamNoSys sequences.

		DEV SET		
Approach:	Supervision	BLEU-4	BLEU-1	ROUGE
T2G2H	✗	**22.06**	**47.55**	**35.74**
T2G2H	✓	21.79	46.21	35.99
T2H	✗	26.14	**44.35**	**50.05**
T2H	✓	**26.99**	42.73	48.85

5. State-of-the-art Comparisons

We achieve a BLEU-4 score of **26.99** on the MeineDGS dataset and **25.09** on PHOENIX14T, two new state-of-the-art baselines.

PHOENIX14T;

	DEV SET		
Approach:	BLEU-4	BLEU-1	ROUGE
T2G (Stoll et al., 2018)	16.34	50.15	48.42
T2G (Saunders et al., 2020)	20.23	55.65	55.41
T2G (Li et al., 2021)	18.89	-	49.91
T2G (Moryossef et al., 2021)	23.17	-	-
T2G Baseline (ours)	22.47	58.98	57.96
T2G Best Model (ours)	**25.09**	**60.04**	**58.82**

MeineDGS;

	DEV SET		
Approach:	BLEU-4	BLEU-1	ROUGE
T2G (Saunders et al., 2022)	3.17	-	32.93
T2G Our best	10.5	33.56	35.79
T2G2H Our best	22.06	47.55	36.20
T2H Our best	**26.99**	**42.73**	**48.89**

Acknowledgements

This project was supported by Adam Munder, Mariam Rahmani and Marina Lovell from OmniBridge, an Intel Venture. We thank the SNSF Sinergia project 'SMILE II' (CRSII5 193686), the European Union's Horizon2020 research project EASIER (101016982) and the EPSRC project 'ExTOL' (EP/R03298X/1). This work reflects only the authors view and the Commission is not responsible for any use that may be made of the information it contains.