Neural Machine Translation Methods for Translating Text to Sign Language Glosses

Dele Zhu[1], Vera Czehmann[1, 2] and Eleftherios Avramidis[2]

1. Technical University of Berlin, Berlin, Germany
2. German Research Center for Artificial Intelligence (DFKI), Berlin, Germany

TECHNISCHE
UNIVERSITÄT
BERLIN

Deutsches
Forschungszentrum
für Künstliche
Intelligenz GmbH

Introduction

Association for Computational Linguistics
ACL 2023

Glosses:

- written representation of Sign Languages (SL)
- limited representation ability, but useful for:
  - interpreters and educational uses
  - intermediate step for spoken to SL translation
  - investigating methods that may apply to more accurate representations in the future

Objective:

- Use low-resource MT techniques successful for spoken languages

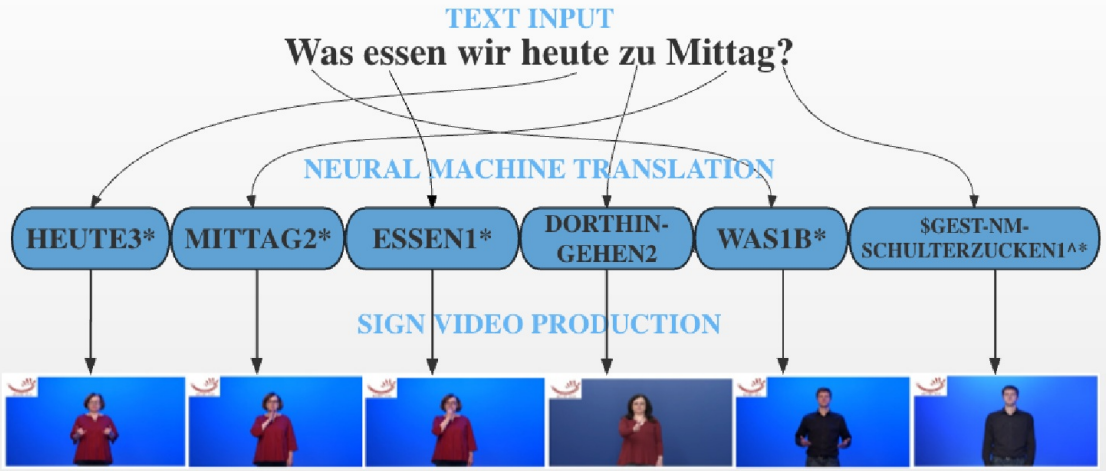


Fig.1 Text-to-video SLT using glosses as an intermediate step (Source of images: Müller et al., 2020).

Approach Overview

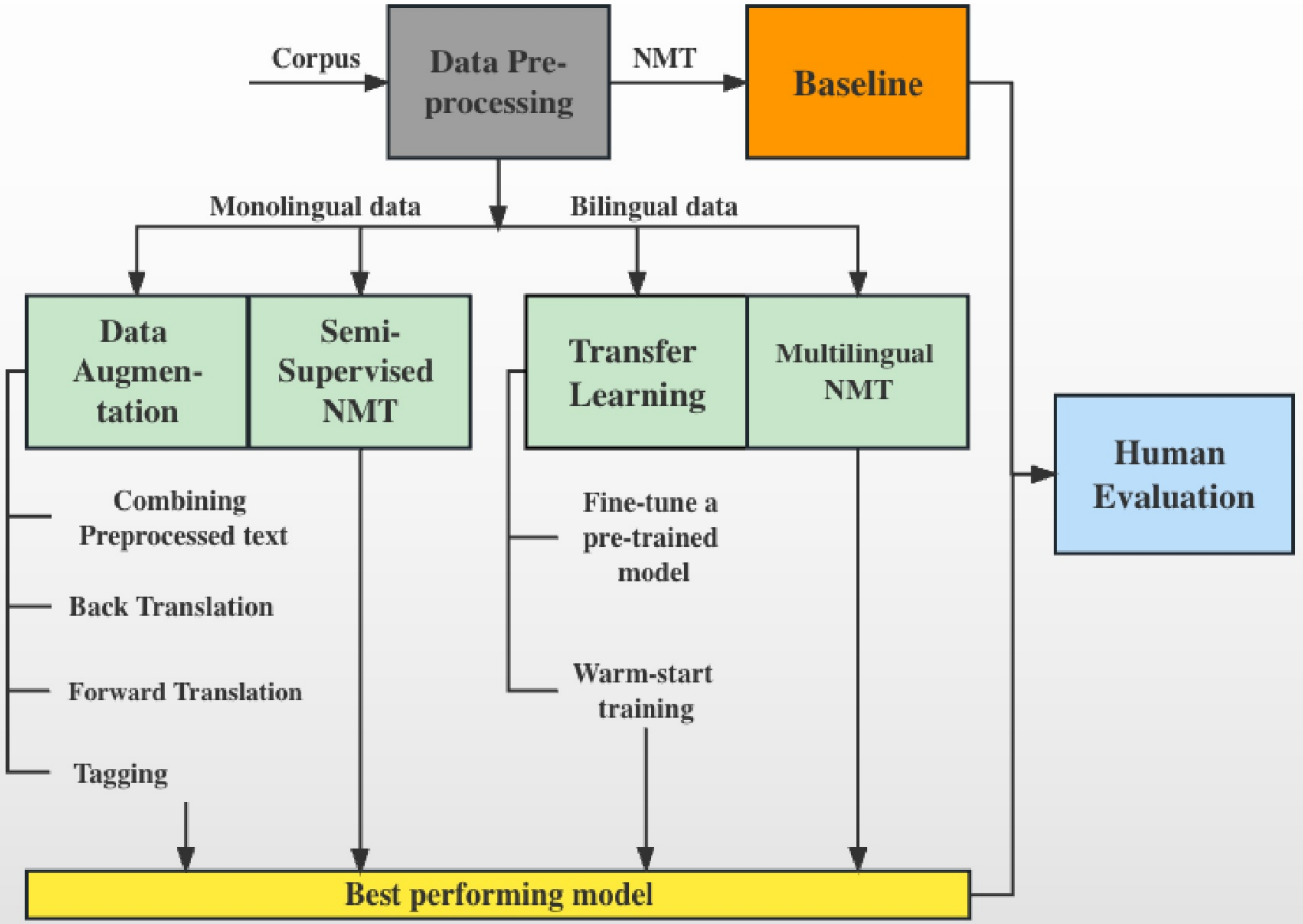


Fig.2 Scheme of experiments

Datasets

Main experiments:

- German sign language - Deutsche Gebärdensprache (DGS)
  1. RWTH-PHOENIX-Weather 2014T (Camgoz et al., 2018)
  2. The Public DGS Corpus (Hanke et al., 2020)

Generalization experiments:

- American sign language (ASL)
  NCSLGR (Vogler and Neidle 2012)

Experiments - Baseline & Data Augmenation & Semi-supervised NMT

Baseline:

- Transformer adapted to LR settings (1 encoder, 2 decoders; Gu et al 2018)

Data Augmentation:

- **Combined preprocessed text:** Collection of source text with different preprocessing techniques
- **Back translation** (Sennrich et al., 2016a): Additional source data translating target data (glosses) from target-to-source model
- **Forward translation** (Zhang and Zong, 2016): Additional synthetic target data translating a source-language monolingual dataset with the baseline system
- **Tagging** (Caswell et al., 2019): Add special prefix before synthetic sentences

Semi-supervised NMT:

- Copy a monoligual dataset to both source & target side (Currey et al. 2017) and combine with the SL parallel dataset

Experiments - Transfer Learning & Multilingual NMT

Transfer Learning:

- **Fine-tuning** of a pre-trained model: Opus-MT de-en model
- **Warm-start training** (Nguyen and Chiang, 2017):
  - Train the parent model with large-scale and SL train set
  - Fine-tune the pre-trained model with SL dataset

Multilingual NMT (Johnson et al., 2017):

- Train NMT system with both SL dataset and large-scale de-en dataset
- Add **target-language-indicator** before each source sentence

German-to-English: **<2en>** Wie heißt du? →What is your name?
German-to-DGSglosses: **<2gloss>** im süden freundliches wetter→sued region besser

Results

Corpus	System	BPE Vocab	Dev BLEU	Dev ChrF	Dev TER	Test BLEU	Test ChrF	Test TER
PHOENIX	Baseline	2k	22.78	51.87	55.84	20.14	52.04	56.12
	Combine	2k	24.01	52.32	53.20	21.88	51.51	54.53
	Semi+Tag	32k	26.55	55.76	50.83	24.15	55.13	51.17
	Fine-tune	65k	26.39	**56.84**	50.88	24.67	**55.97**	52.86
	Warm	32k	**27.62**	**56.92**	**49.25**	24.89	55.46	**50.40**
	Multi+back	32k	**28.41**	**57.54**	49.39	**26.32**	**56.70**	51.15
	Multi-big+back	32k	**28.53**	**57.64**	**48.93**	**25.98**	**56.67**	50.94
DGS	Baseline	5k	4.04	31.20	79.34	3.13	30.38	78.64
	Combine	5k	3.71	29.97	80.21	2.75	29.31	80.01
	Semi+Tag	32k	5.00	32.69	79.47	4.10	31.30	78.67
	Fine-tune	65k	5.82	35.05	79.92	4.53	**34.14**	78.98
	Warm	32k	5.87	33.42	**74.07**	4.55	31.90	74.54
	Multi+back	32k	5.35	33.43	78.30	4.85	32.16	76.76
	Multi-big+back	32k	**6.82**	**35.57**	76.37	**5.78**	**33.87**	76.12

Table 1: Selected automatic metric scores of extensive experimentation search of the two DGS corpora. We **boldface** all the values that are not statistically significantly different from the best value of each evaluation metric and underline the results that are statistically significantly higher than baseline at the 95% confidence level.

Comparison with previous work – generalization test on ASL

Approach	Dev BLEU↑	Test BLEU↑
Amin et al. (2021)	-	10.42
Egea Gómez et al. (2021)†	-	13.13
Stoll et al. (2020)	16.34	15.26
Zhang and Duh (2021)	-	16.43
Li et al. (2021)	-	18.89
Saunders et al. (2020b)	20.23	19.10
Saunders et al. (2022)	21.93	20.08
Egea Gómez et al. (2022)	-	20.57
Walsh et al. (2022)	25.09	23.19
Our PHOENIX Multi+back	**28.41**	**26.32**

Table 4: Results comparison with recent work. (†) We compute the BLEU by ourselves, as the authors of paper only present the BLEU score in character level.

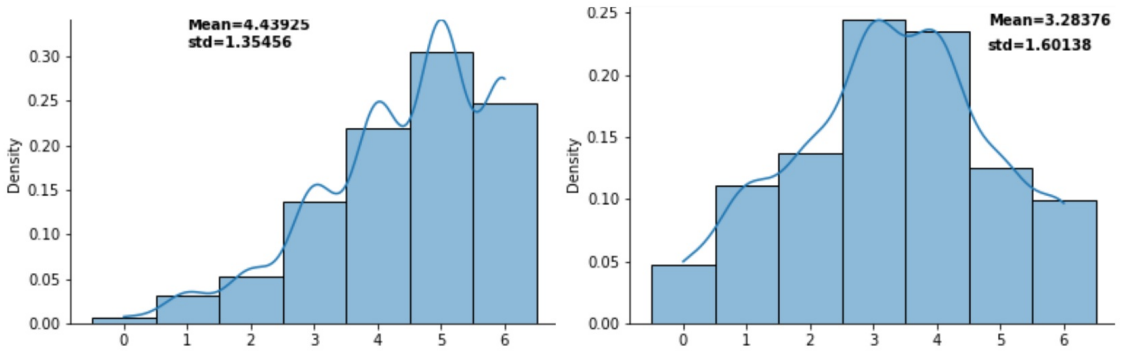
System	Test BLEU	Test ChrF	Test TER
Baseline	10.50	30.65	**78.95**
Back	9.37	28.25	**78.67**
Warm-start	**12.11**	33.53	83.44
Multi	**12.35**	**38.33**	**78.26**

Table 2: Automatic metric scores for NCSLGR corpus.

Human Evaluation

System	Size	automatic			human	
		BLEU↑	ChrF↑	TER↑	Mean↑	Std↑
PHOENIX Egea Gómez et al. (2021)	642	13.13	46.86	73.33	2.74	1.64
PHOENIX baseline		20.14	52.04	56.12	3.85	1.58
PHOENIX Multi+back		**26.32**	**56.70**	**51.15**	**4.44**	**1.35**
DGS Baseline (sampled 10%)	511	3.44	29.56	78.55	2.49	1.81
DGS Multi-big (sampled 10%)		**6.97**	**33.16**	**73.45**	**3.28**	**1.60**

Table 3: System comparison based on the human evaluation. The **bold-faced** systems are significantly better than the respective baselines.



(a) PHOENIX Multi+back (b) DGS Multi-big

Figure 3: Density of human evaluation scores for the two best-scoring systems.

Conclusion & Contributions

this paper is the first work on text-to-gloss machine translation

- to achieve significant improvements, as compared to the baseline methods and related work, on the two known natural German SL datasets annotated with glosses,

- to perform extensive experimentation with most known LRL-related MT methods and their combinations and in particular:

  - to apply **semi-supervised NMT** by copying the monolingual data to both the source and target side
  - to use **transfer learning** via the warm-start strategy, and
  - to use a **multilingual NMT** setting with the focus on improving the text-to-gloss direction.

Thanks for your attention!