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
2. German Research Center for Artificial Intelligence (DFKI)

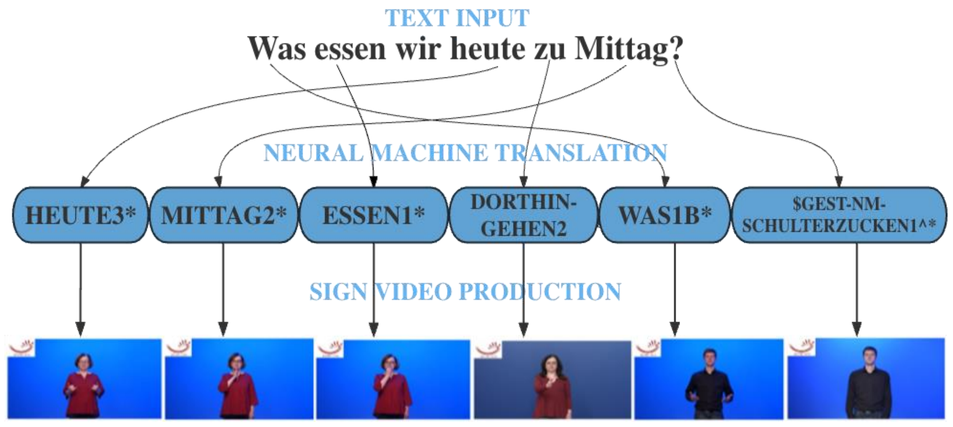
Introduction

Glosses

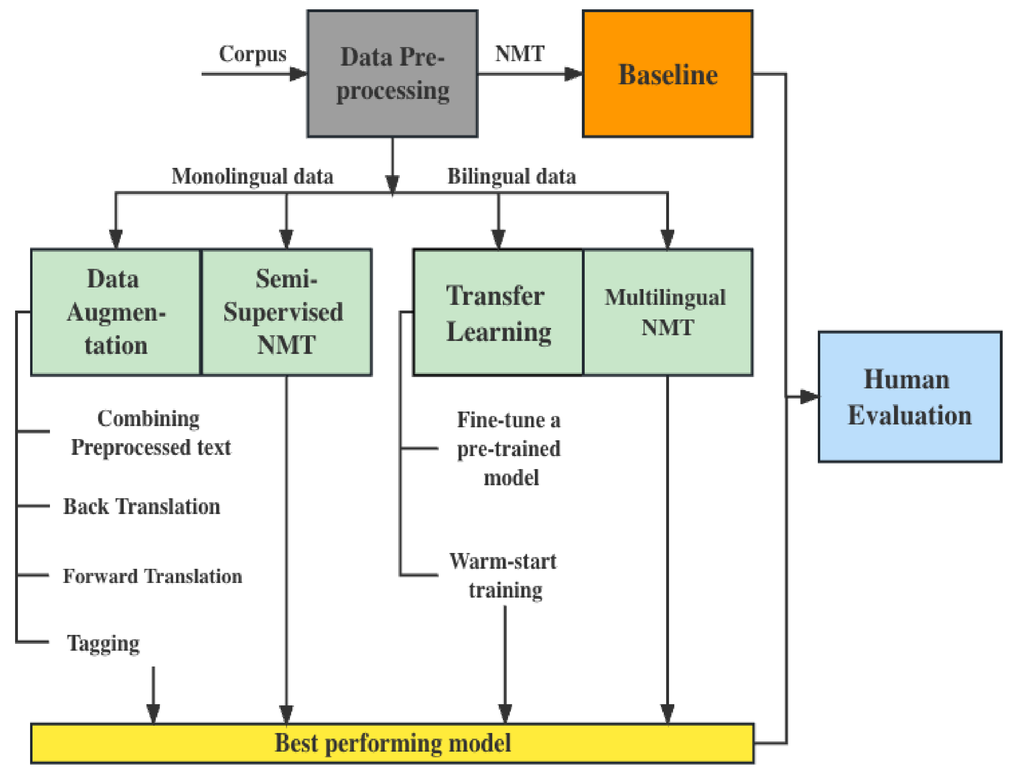
- Written representation of Sign Languages (SL)
- Limited representation ability, but useful for
  - Interpreters and educational uses
  - Intermediate step for spoken to SL translation
  - Investigating methods that may apply to more accurate representations in the future

Objective

- Use techniques successful in low-resource spoken languages MT to improve Sign Language Translation (SLT).



Approach overview



Methods

Data augmentation

- Combined preprocessed text, Back-translation, Forward-translation and Tagging

Semi-supervised NMT

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

Transfer learning

- Fine-Tuning of a pre-trained model: Opus-MT de-en model
- Warm-start training (Nguyen and Chiang, 2007)

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

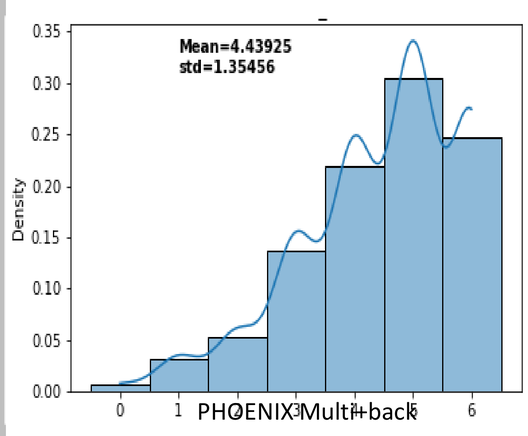
Results

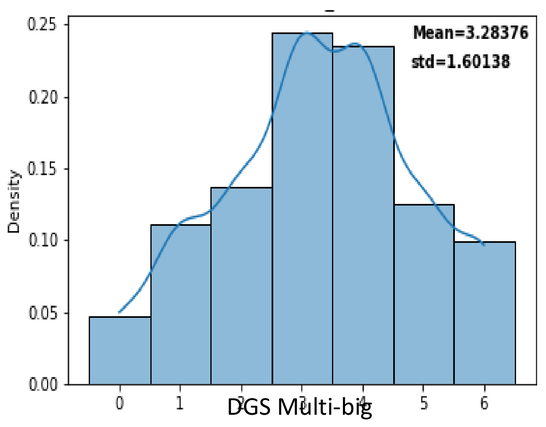
Automatic MT metrics results for possible settings

	System	BPE Vocab	Dev BLEU	Dev ChrF	Dev TER	Test BLEU	Test ChrF	Test TER
RWTH Phoenix-2014T	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
	Combine+Tag	2k	22.94	52.09	52.88	21.11	51.65	54.81
	Back	2k	23.63	52.03	53.98	21.04	51.59	54.57
	Back+Tag	2k	23.62	52.85	52.88	21.57	52.41	53.94
	Forward	2k	23.03	52.56	53.71	20.40	51.54	55.63
	Forward+Tag	2k	23.45	52.49	54.16	21.64	52.27	54.57
	All_combined	2k	23.63	52.32	54.19	21.04	51.97	54.71
	Semi	32k	26.76	55.41	51.10	22.67	53.87	53.07
	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	32k	**28.34**	**57.29**	**48.48**	24.30	55.71	51.03
	Multi-big	32k	**27.45**	56.52	**48.77**	**24.97**	55.75	**49.89**
	Multi+combine	32k	26.61	55.59	50.21	23.22	54.55	52.84
	Multi-big+combine	32k	**28.02**	**57.07**	**49.31**	24.94	55.89	51.01
	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
Public DGS Corpus	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
	Combine+Tag	5k	3.23	28.69	81.31	2.27	28.17	81.03
	Back	5k	3.83	30.08	82.75	3.06	29.30	80.94
	Back+Tag	5k	3.88	29.66	79.55	2.75	28.91	79.05
	Forward	5k	3.51	29.14	83.03	2.81	28.24	81.13
	Forward+Tag	5k	3.75	29.69	86.20	2.93	29.06	83.21
	All_combine	5k	3.14	28.37	81.61	2.43	27.87	81.83
	Semi	32k	5.16	33.43	76.19	4.42	31.81	76.35
	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	32k	6.06	35.18	74.51	5.32	33.55	74.71
	Multi-big	32k	**6.60**	**35.26**	**73.25**	**5.46**	33.49	**73.53**
	Multi+combine	32k	4.64	32.33	80.39	3.85	31.38	78.34
	Multi-big+combine	32k	**6.79**	**35.50**	**73.98**	**5.61**	**33.88**	**73.94**
	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

Human evaluation for best scoring systems

System	Size	automatic BLEU↑	automatic ChrF↑	automatic TER↑	human Mean↑	human 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**





Comparison with previous work

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**

Conclusion

The first work on text-to-gloss machine translation

- to achieve significant improvements on the two known German SL datasets annotated with glosses
- to perform extensive experimentation with most known LRL-related MT methods and their combinations, in particular: **semi-supervised NMT, transfer learning via warm-start and Multilingual NMT**

This work was supported by the German Federal Ministry of Education and Research (BMBF) through the project SocialWear (grant no. 01IW20002)