Applying Lessons from Low-Resource Machine Translation to Speech and Sign Language Translation

Rico Sennrich

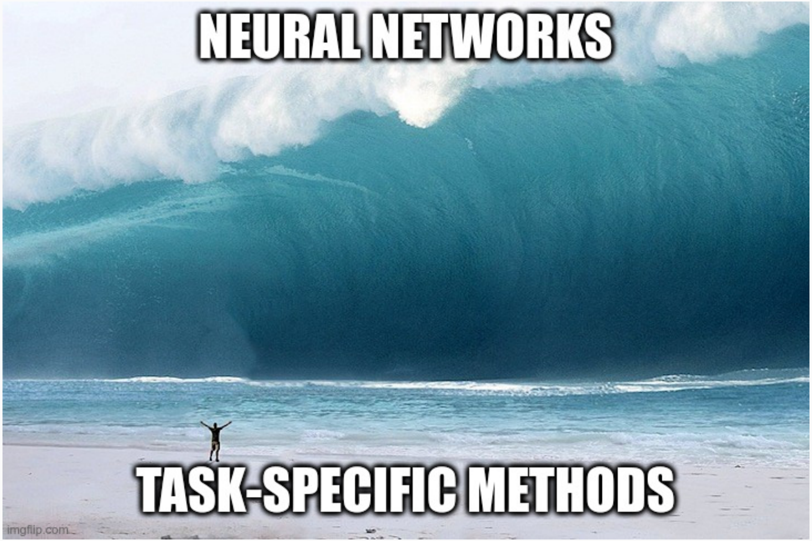
May 6 2023

University of Zurich
University of Edinburgh

1. Natural Language Processing for Text
   - Lessons from Low-Resource MT
2. Transfer in Multimodal Tasks
   - Speech Translation
   - Sign Language Translation

Natural Language Processing for Text

2 Waves of Unification in Natural Language Processing



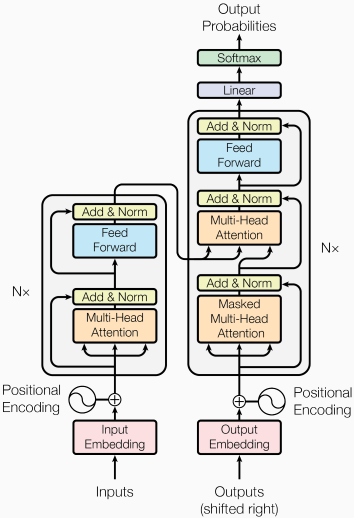
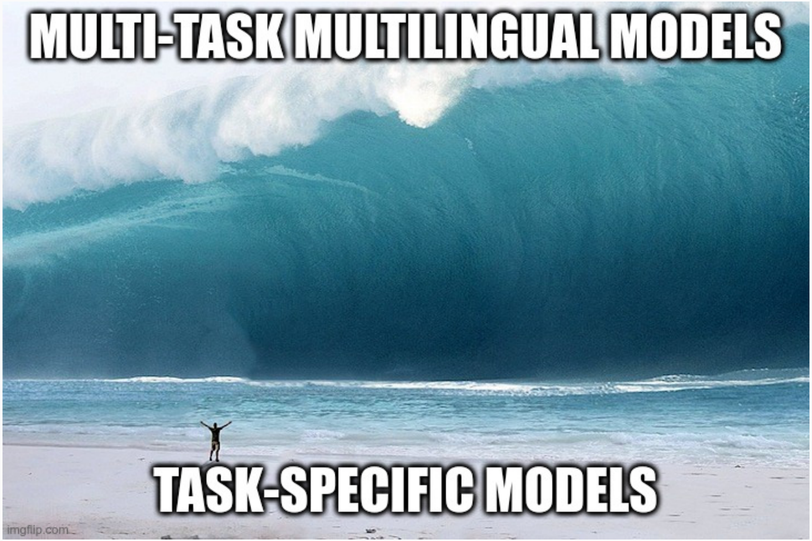


Figure 1: The Transformer - model architecture.

2 Waves of Unification in Natural Language Processing



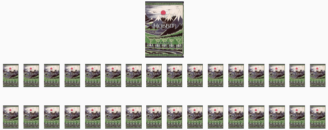
2 Waves of Unification in Natural Language Processing

1. Natural Language Processing for Text
   - Lessons from Low-Resource MT
2. Transfer in Multimodal Tasks
   - Speech Translation
   - Sign Language Translation

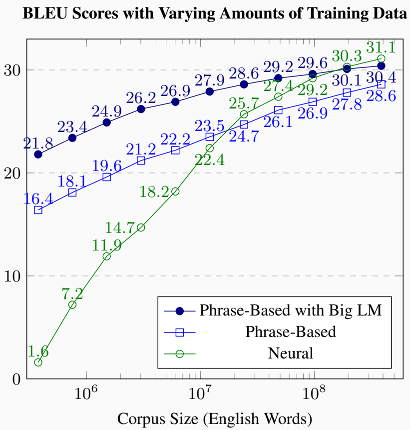
On Data Scale

some EN-DE parallel corpora:

	words	sentences	The Hobbit
The Hobbit	100k	5000	1
TED talks	3.2M	160000	32
Europarl	50M	2M	500
Opensubtitles	170M	14M	1700
Paracrawl	4300M	280M	43 000



2016–2017: Painful Start for Neural MT

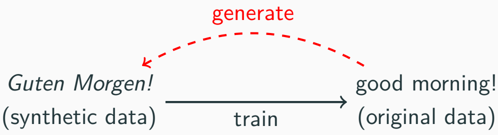


(English→Spanish)

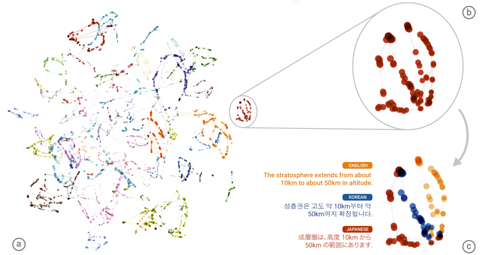
[Koehn and Knowles, 2017]

2015–2018: Monolingual and Multilingual Data

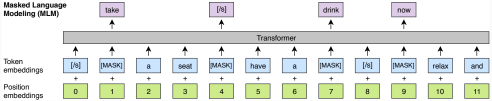
backtranslation

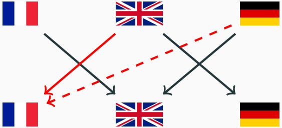


multilingual models



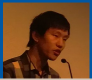
self-supervised pre-training

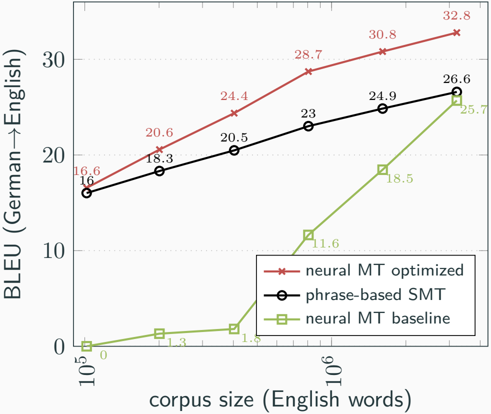




2019: With Right Hyperparameters, You May Not Need Extra Data

[Sennrich and Zhang, ACL 2019]

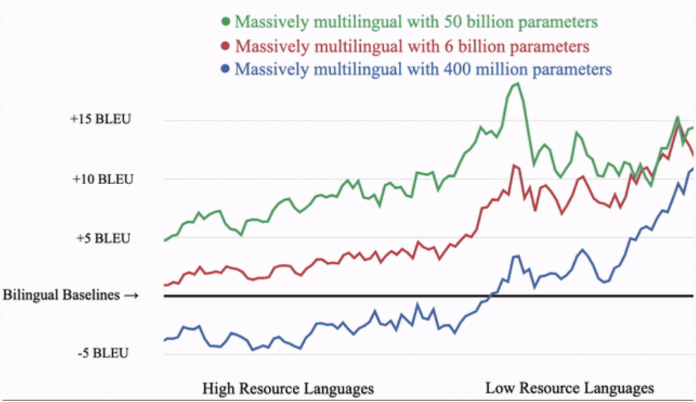




biggest improvements:

- widely-used innovations already help (tied embeddings, layer normalization, label smoothing...)
- tune subword vocabulary size
- apply aggressive regularization (dropout)

Today: Large, Massively Multilingual Models



https://ai.googleblog.com/2019/10/exploring-massively-multilingual.html

Transfer in Multimodal Tasks

Transfer in Multimodal Tasks

Speech Translation

First Wave: Unification of Architectures

automatic speech recognition

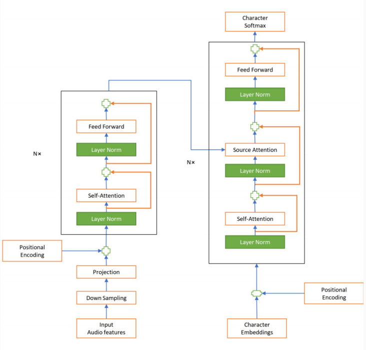


Figure 1: *Our ASR Transformer-based Architecture.*

[Pham et al., 2019]

sign language translation

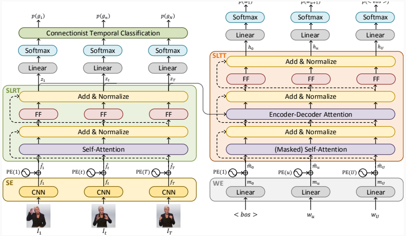


Figure 2: A detailed overview of a single layered Sign Language Transformer.
(SE: Spatial Embedding, WE: Word Embedding , PE: Positional Encoding, FF: Feed Forward)

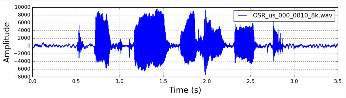
[Camgoz et al., 2020]

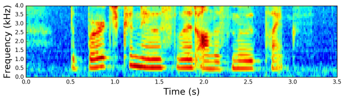
Adapting Transformer to Multimodal Data

simple principle:

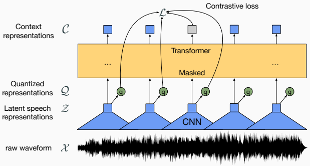
we can input speech/visual data as vectors instead of word embeddings

raw waveforms	extracted features (e.g. log-mel filterbanks)	self-supervised embeddings (e.g. wav2vec)





[Fayek, 2016]



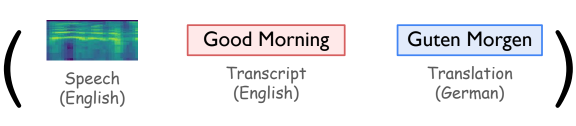
[Baevski et al., 2020]

Training Data for End-to-End Speech Translation

corpus	language pairs	domain	segments
Fisher [Post et al., 2013]	ES→EN	telephone	140k
LibriTrans [Kocabiyikoglu et al., 2018]	EN→FR	audiobooks	130k
MuST-C [Di Gangi et al., 2019]	EN→{DE,ES,FR,IT,NL,PT,RO,RU}	TED talks	250k

→ low-resource scenario

...but we typically have training triplets:



Transfer Learning for End-to-End Speech Translation Systems

common solutions: auxiliary tasks in addition to end-to-end model $P(T|X)$ [Weiss et al., 2017, Bérard et al., 2018]:

- parameter sharing with ASR system: $P(S|X)$
- parameter sharing with MT system: $P(T|S)$

less common: synthetic training data [Jia et al., 2018]

- use ASR data; create target side via MT
- use MT data; create speech input via text-to-speech (TTS)

A Word on wav2vec

claim [Baevski et al., 2020]

For ASR, ”using just ten minutes of labeled data [...] achieves 4.8[%] WER“

counter [San et al., 2023]

unrealistic for actual low-resource languages:

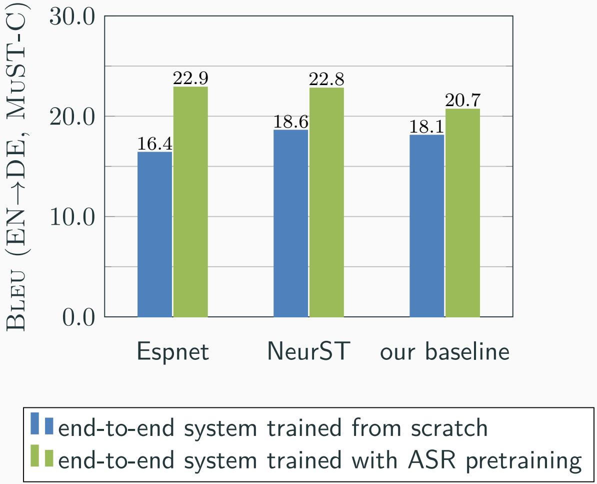
- relies on large language model (803M tokens; English)

without LM	40% WER
with realistically-sized LM (80k tokens)	24% WER

- relies on similarity between pre-training and test languages (results without LM)

on Gronings and Frisian (Germanic)	44-53% WER
on Besemah and Nasal (Malayo-Polynesian)	62-70% WER

Transfer Learning for End-to-End Spoken Language Translation Systems



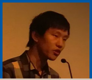
Problems with Reliance on Transcripts

Transcripts are not always available
$\rightarrow$ many languages have no written form

Questioning assumptions for its own sake
$\rightarrow$ focus on transfer learning may detract from other considerations

Revisiting End-to-End Speech Translation from Scratch

[Zhang, Haddow, Sennrich, ICML 2022]



CTC Regularization

- Use translation as CTC labels
- No transcripts are used

Parameterized Distance Penalty (PDP)

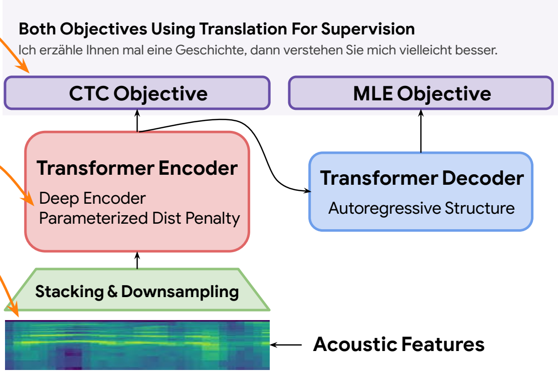
- Add freedom in local attention modeling

Neural Acoustic Feature Modeling

- Use raw waveform to retain local details

Hyperparameter Tuning

- Beam search; Model depth/width



Using speech-translation pairs alone **with no transcripts**

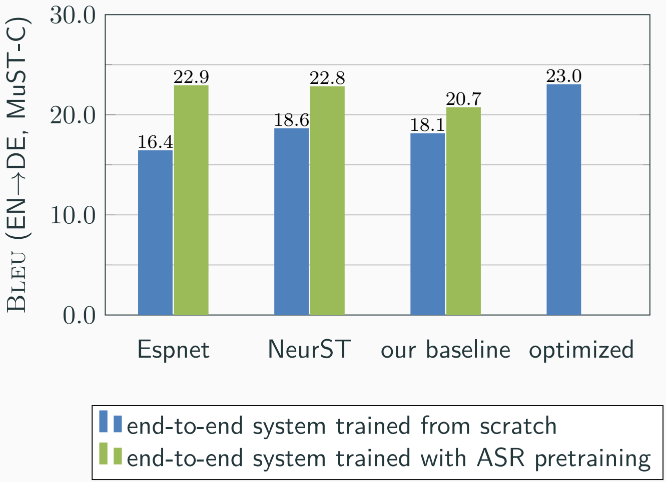
Revisiting End-to-End Speech Translation from Scratch: Results

System	BLEU		Avg
NeurST (pretrain-finetune)	22.8		24.9
Baseline	18.1		-
+ hyperparameter tuning	21.1	+3.0	-
+ PDP (R=512)	21.8	+0.7	-
+ CTC regularization	22.7	+0.9	-
+ neural acoustic model	**23.0**	+0.3	**25.2**

Test performance on MuST-C En-De and average results on the other language pairs

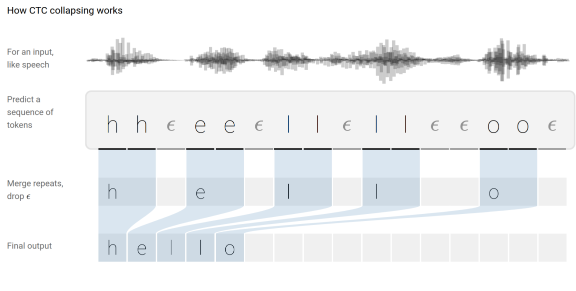
Note all our models are trained with speech-translation pairs alone

Revisiting End-to-End Speech Translation from Scratch: Results



More on CTC Regularization

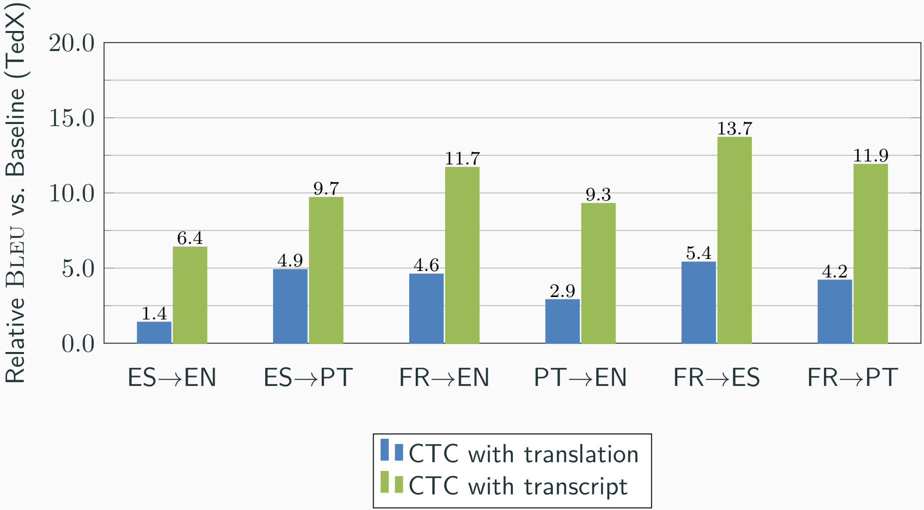
originally developed for monotonic tasks without alignment (handwriting, ASR)



relatively recent finding that CTC also helps with translation as labels:

- non-autoregressive machine translation [Libovický and Helcl, 2018]
- autoregressive spoken language translation [Zhang et al., 2022]
- autoregressive machine translation [Yan et al., 2022]

CTC Loss: Transcripts Still Useful if Available



results from [Zhang et al., 2023a]

Transfer in Multimodal Tasks

Sign Language Translation

On Sign Languages

Why Sign Language Translation?

- sign language is not universal
  $\rightarrow$ several hundred sign languages worldwide
- "spoken" languages are foreign languages
  $\rightarrow$ no linguistic relation between German Sign Language and German

Why Sign Language Translation?

- sign language is not universal
  $\rightarrow$ several hundred sign languages worldwide
- "spoken" languages are foreign languages
  $\rightarrow$ no linguistic relation between German Sign Language and German

sign language research at University of Zurich

Similarities and Differences to Other Tasks

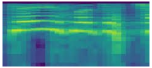
common with low-resource translation:
training data is sparse, but potential for cross-lingual and cross-task transfer

common with speech translation:
input modality (audio/video) is barrier for transfer

common with some spoken languages:
sign languages have no commonly used and closely aligned “written form”

Sign Language: Representations

spoken languages:

( Speech (English) | **Good Morning** Transcript (English) | **Guten Morgen** Translation (German))

sign languages:

Video (German Sign Language)

?

Guten Morgen

Translation (German)

Sign Language: Representations

Video
(German Sign Language)

?

Guten Morgen

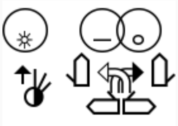
Translation
(German)

Sign Language: Representations

Video (German Sign Language)	GUTEN MORGEN Glosses	Guten Morgen Translation (German)

Sign Language: Representations

Video
(German Sign Language)



SignWriting

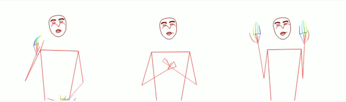
Guten Morgen

Translation
(German)

Sign Language: Representations

Video

(German Sign Language)



Poses

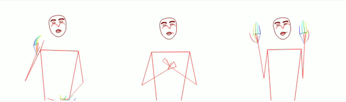
Guten Morgen

Translation

(German)

Sign Language: Representations

Video
(German Sign Language)



Poses

Guten Morgen

Translation
(German)

intermediate representations:

- could help end-to-end system as auxiliary task
- could help cascade systems (lower-dimensional than video)
- unlike speech translation, there is no extra data “for free” (ASR/MT)

Multi-Task Training for Sign Language Translation

[Zhang, Müller, Sennrich, ICLR 2023]

goals:

- build optimzed sign language translation system
- measure benefits of multi-task training (using glosses and MT)
- test sign language translation on more challenging dataset

Sign Language Datasets with Glosses

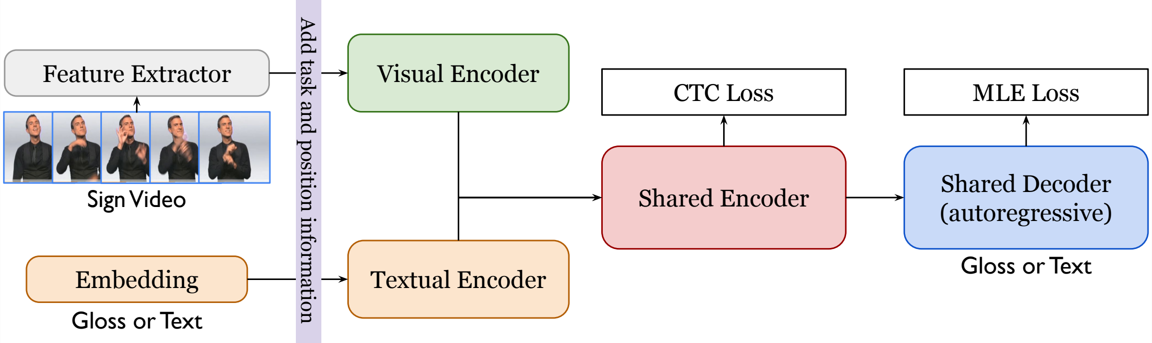
	Phoenix-2014T [Camgoz et al., 2018]	CSL-Daily [Zhou et al., 2021]	DGS3 [Hanke et al., 2020]
signers	9	10	330
glosses	1085	2000	8580
domain	weather	daily life	diverse
train segments	7096	18401	60306
source	German Sign Language	Chinese Sign Language	German Sign Language
target	German	Chinese	German

Phoenix-2014T and CSL-Daily dominate previous work

we are first to attempt end-to-end sign language translation on DGS3

side remark: [Zhou et al., 2021]: Improving Sign Language Translation with Monolingual Data by Sign Back-Translation

SLTUnet Architecture



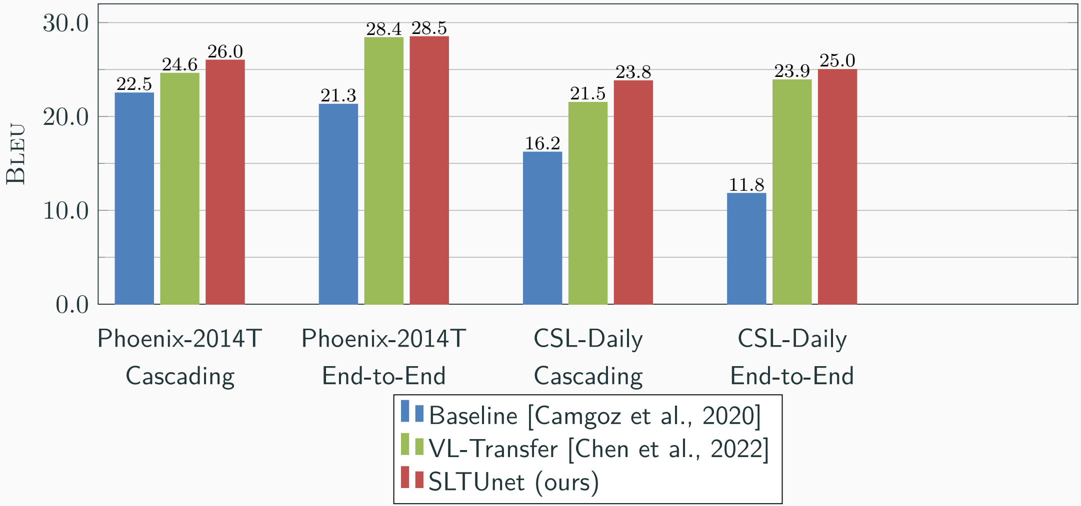
Task	Task Tag	Input	Output	Training Objective
Sign2Gloss	*[2gls]*	sign video	gloss	$\alpha\mathcal{L}^{\text{CTC}}(\text{gloss}) + \mathcal{L}^{\text{MLE}}(\text{gloss})$
Sign2Text	*[2txt]*	sign video	text	$\alpha\mathcal{L}^{\text{CTC}}(\text{gloss}) + \mathcal{L}^{\text{MLE}}(\text{text})$
Gloss2Text	*[2txt]*	gloss	text	$\mathcal{L}^{\text{MLE}}(\text{text})$
Text2Gloss	*[2gls]*	text	gloss	$\mathcal{L}^{\text{MLE}}(\text{gloss})$
Text2Text (MT)	*[2txt]*	source text	target text	$\mathcal{L}^{\text{MLE}}(\text{target})$

Some Insights from Optimization

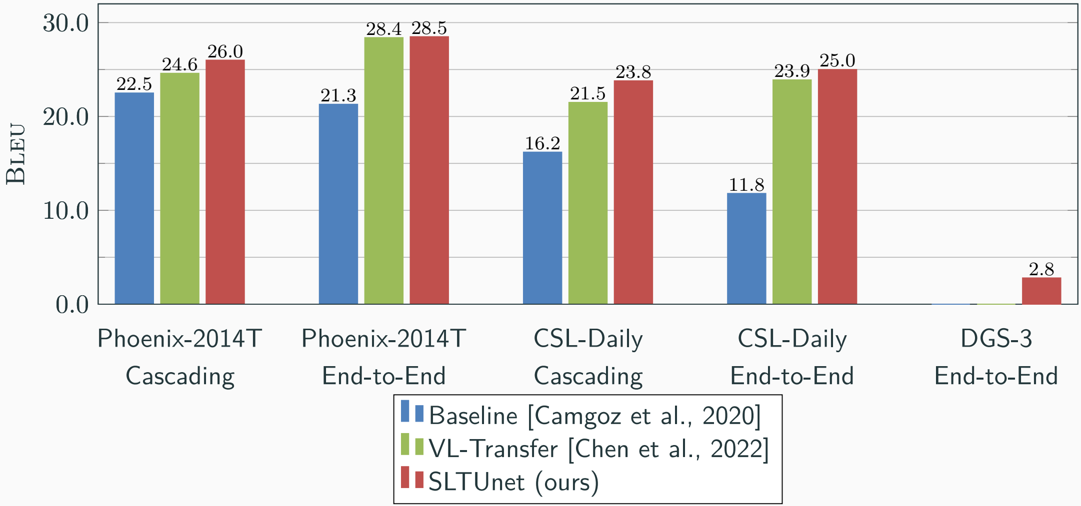
biggest improvements over our baseline (PHOENIX-2014T dev):

- CTC regularization (+2.8 BLEU)
- BPE dropout (50%) (+1 BLEU)
- multi-task training (+1 BLEU)
  $\rightarrow$ but extra MT data only helps little (+0.1 BLEU)

Sign Language Translation Results



Sign Language Translation Results



DGS-3 Error Analysis

- is this too hard for end-to-end modelling?
  not only that: cascade similarly fails
- for cascade, where does it fail?
  sign→gloss has WER of 67% (!)
- model shows hints of translation, but majority is hallucinated:

Gold Gloss:	MORGEN3 FISCH1 MARKT4 BEKANNT1 $INDEX2
Gold Text:	Morgens geht man zum Fischmarkt, der ist bekannt. (*In the morning you go to the fish market, it's well known.*)
SLTUnet	Ja, das ist bekannt. (*Yes, that is known.*)

Summary

- LoResMT community can (and should) contribute to modalities beyond text
  - we can apply our expertise successfully
  - interesting challenges to be solved
  - many less-privileged languages are not text-based
- knowledge sharing (cross-lingual; cross-task) is workhorse for low-resource MT...
  ...but it's encouraging how far we can get with regularization and little data

Thank you for your attention

Resources

- code for speech translation:
  `https://github.com/bzhangGo/zero`
- code for sign language translation:
  `https://github.com/bzhangGo/sltunet`

this research was funded by Swiss National Science Foundation

European Commission

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020).
wav2vec 2.0: A framework for self-supervised learning of speech representations.
In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, Advances in Neural Information Processing Systems, volume 33, pages 12449–12460. Curran Associates, Inc.

Bérard, A., Besacier, L., Kocabiyikoglu, A. C., and Pietquin, O. (2018).
End-to-End Automatic Speech Translation of Audiobooks.
In
ICASSP 2018 - IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, Alberta, Canada.

Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., and Bowden, R. (2018).
Neural sign language translation.
In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Camgoz, N. C., Koller, O., Hadfield, S., and Bowden, R. (2020).
Sign language transformers: Joint end-to-end sign language recognition and translation.
In IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Chen, Y., Wei, F., Sun, X., Wu, Z., and Lin, S. (2022).
A simple multi-modality transfer learning baseline for sign language translation.
In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 5120–5130.

CONNEAU, A. and Lample, G. (2019).
Cross-lingual language model pretraining.
In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc.

Di Gangi, M. A., Cattoni, R., Bentivogli, L., Negri, M., and Turchi, M. (2019).
MuST-C: a Multilingual Speech Translation Corpus.
In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Fayek, H. M. (2016).
Speech processing for machine learning: Filter banks, mel-frequency cepstral coefficients (mfccs) and what's in-between.

Hanke, T., König, S., Konrad, R., Langer, G., Barbeito Rey-Geißler, P., Blanck, D., Goldschmidt, S., Hofmann, I., Hong, S.-E., Jeziorski, O., Kleyboldt, T., König, L., Matthes, S., Nishio, R., Rathmann, C., Salden, U., Wagner, S., and Worseck, S. (2020). **MEINE DGS. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release.**

Jia, Y., Johnson, M., Macherey, W., Weiss, R. J., Cao, Y., Chiu, C., Ari, N., Laurenzo, S., and Wu, Y. (2018).
Leveraging weakly supervised data to improve end-to-end speech-to-text translation.
CoRR, abs/1811.02050.

Kocabiyikoglu, A. C., Besacier, L., and Kraif, O. (2018).
Augmenting librispeech with French translations: A multimodal corpus for direct speech translation evaluation.
In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Koehn, P. and Knowles, R. (2017).
Six Challenges for Neural Machine Translation.
In Proceedings of the First Workshop on Neural Machine Translation, pages 28–39, Vancouver.

Libovický, J. and Helcl, J. (2018).
End-to-end non-autoregressive neural machine translation with connectionist temporal classification.
In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3016–3021, Brussels, Belgium. Association for Computational Linguistics.

Pham, N.-Q., Nguyen, T.-S., Ha, T.-L., Hussain, J., Schneider, F., Niehues, J., Stüker, S., and Waibel, A. (2019).
The iwslt 2019 kit speech translation system.
In International Workshop on Spoken Language Translation.

Post, M., Kumar, G., Lopez, A., Karakos, D., Callison-Burch, C., and Khudanpur, S. (2013).
Improved speech-to-text translation with the fisher and callhome spanish–english speech translation corpus.
In IWSLT13.

San, N., Bartelds, M., Billings, B., de Falco, E., Feriza, H., Safri, J., Sahrozi, W., Foley, B., McDonnell, B., and Jurafsky, D. (2023).
Leveraging supplementary text data to kick-start automatic speech recognition system development with limited transcriptions.
In Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 1–6, Remote. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016).
Improving neural machine translation models with monolingual data.
In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R. and Zhang, B. (2019).
Revisiting low-resource neural machine translation: A case study.
In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017).
Attention is All you Need.
In Advances in Neural Information Processing Systems 30, pages 5998–6008.

Weiss, R. J., Chorowski, J., Jaitly, N., Wu, Y., and Chen, Z. (2017).
Sequence-to-Sequence Models Can Directly Translate Foreign Speech.

Yan, B., Dalmia, S., Higuchi, Y., Neubig, G., Metze, F., Black, A. W., and Watanabe, S. (2022).
Ctc alignments improve autoregressive translation.

Zhang, B., Haddow, B., and Sennrich, R. (2022).
Revisiting end-to-end speech-to-text translation from scratch.
In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pages 26193–26205. PMLR.

Zhang, B., Haddow, B., and Sennrich, R. (2023a).
Efficient ctc regularization via coarse labels for end-to-end speech translation.
In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Dubrovnik, Croatia.

Zhang, B., Müller, M., and Sennrich, R. (2023b).
SLTUNET: A simple unified model for sign language translation.
In The Eleventh International Conference on Learning Representations, Kigali, Rwanda.

Zhou, H., Zhou, W., Qi, W., Pu, J., and Li, H. (2021).
Improving sign language translation with monolingual data by sign back-translation.
In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1316–1325.