Neural Sign Language Translation

Necati Cihan Camgoz **Simon Hadfield** **Oscar Koller** **Hermann Ney** **Richard Bowden**

Centre for Vision Speech and Signal Processing, Faculty of Engineering and Physical Sciences, University Of Surrey, Guildford, Surrey, GU2 7XH {n.camgoz, s.hadfield, r.bowden}@surrey.ac.uk

Human Language Technology and Pattern Recognition Group, RWTH-Aachen University, Aachen, Germany {koller, ney}@cs.rwth-aachen.de

Sign Language Translation

- ***Recognition ≠ Translation***
- State-of-the-art in ***Recognition***: Continuous Sign Language Recognition
  - Additional step is required to go from sign **glosses** to written **text**.
- Previous attempts on ***Translation***:
  - [1] Recognize isolated signs -> combine them using language model.
  - [2] Use ground truth sign glosses -> train statistical translation models.

Neural Sign Language Translation

- Approach Sign Language Translation as a **machine translation** task.
- **First parallel corpora** for continuous sign videos to spoken language text.
- **First end-to-end continuous sign language translation** (video -> text).
- Variety of baseline results to underpin future research in the field:
  - **Gloss2Text**: Simulate perfect recognition.
  - **Sign2Text**: Going from video without intermediate representations.
  - **Sign2Gloss -> Gloss2Text**: Predict glosses and use the best Gloss2Text.
  - **Sign2Gloss2Text**: Retrain Gloss2Text from scratch.

RWTH-PHOENIX 2014T

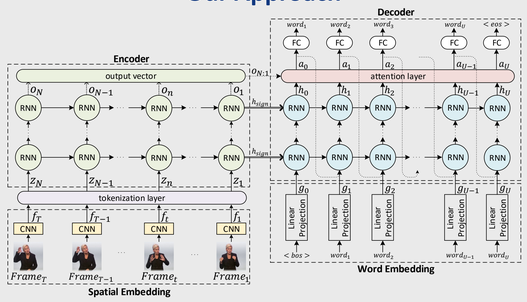
- German Sign Language Videos of Broadcast **Weather Forecasts**.
  - **Extended** version of [3].
- Annotated in **sequence level** for both **sign glosses** and **German text**.
- Sequences realigned to match the spoken language sentences.

	Sign Gloss			German		
	Train	Dev	Test	Train	Dev	Test
Segments	7,096	519	642	← Same		
Full Frames	827,354	55,775	64,627	← Same		
Vocabulary	1,066	393	411	2,887	951	1,001
Total # Tokens	67,781	3,745	4,257	99,081	6,820	7,816
Total # OOV	-	19	22	-	57	60
Singletons	337	-	-	1,077	-	-

Dataset is Available Online

https://www-i6.informatik.rwth-aachen.de/~koller/RWTH-PHOENIX-2014-T/

Our Approach



Spatial Embedding

- **Convolutional Neural Networks**
- **AlexNet** initialized with ImageNet

Tokenization Layer

- Sign2Text: **Identity**
- Sign2Gloss2Text: **BLSTM + FC [4]**

Word Embedding

- **Linear Projection** (Dense)
- Trained from scratch

Encoder-Decoder

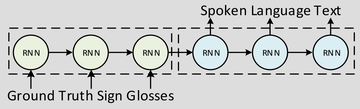
- Units: **GRU** or **LSTM**
- Attention: **Luong** or **Bahdanau**

Quantitative Results

Gloss2Text

What would be the translation performance with perfect recognition?

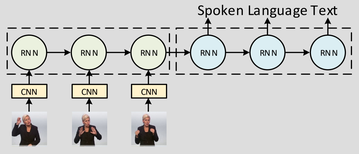
- Virtual upper performance bound
- Encoder-Decoder with GRU and Luong Attention



Sign2Text

Can we go directly from video to text?

- No intermediate representation
- AlexNet fc7 fed to Encoder-Decoder with GRU and Luong Attention



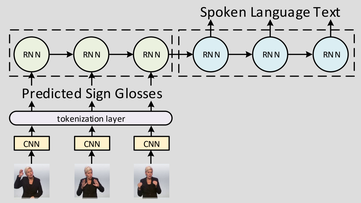
Best Performing	Dev Set			Test Set		
	ROUGE	BLUE-1	BLUE-4	ROUGE	BLEU-1	BLUE-4
Gloss2Text	46.02	44.40	20.16	45.45	44.13	19.26
Sign2Text	31.80	31.87	9.94	31.80	32.24	9.58

Quantitative Results - Cont.

Gloss Level Supervision

Would having gloss level intermediate representations help?

- **Sign2Gloss -> Gloss2Text:** Glosses from [4] fed to best Gloss2Text.
- **Sign2Gloss2Text:** Network trained from scratch using [4]'s predictions .



Best Performing	Dev Set			Test Set		
	ROUGE	BLUE-1	BLUE-4	ROUGE	BLEU-1	BLUE-4
Gloss2Text	46.02	44.40	20.16	45.45	44.13	19.26
Sign2Text	31.80	31.87	9.94	31.80	32.24	9.58
Sign2Gloss -> Gloss2Text	**43.76**	**41.08**	**17.86**	**43.45**	**41.54**	**17.79**
Sign2Gloss2Text	**44.14**	**42.88**	**18.40**	**43.80**	**43.29**	**18.13**

Qualitative Results - Failed Translations

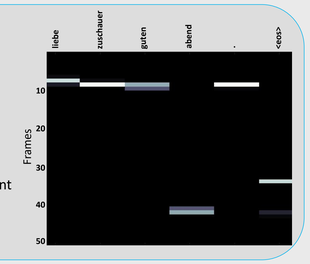
Ground Truth:
DE: und nun die wettervorhersage für morgen **samstag** den **zweiten april** .
*EN: and now the weatherforecast for tomorrow **saturday** the **second april** .*
Gloss2Text:
DE: und nun die wettervorhersage für morgen samstag den **elften** april .
*EN: and now the weatherforecast for tomorrow saturday the **eleventh** april .*
Sign2Text:
DE: und nun die wettervorhersage für morgen **freitag** den **sechsundzwanzigsten märz** .
*EN: and now the weatherforecast for tomorrow **friday** the **twentysixth march** .*
Sign2Gloss2Text:
DE: und nun die wettervorhersage für morgen samstag den **siebzehnten** april .
*EN: and now the weatherforecast for tomorrow saturday the **seventeenth** april .*

Ground Truth:
DE: **die neue woche beginnt** wechselhaft und **kühler** .
*EN: **the new week starts** unpredictable and **cooler** .*
Gloss2Text:
DE: die neue woche beginnt wechselhaft und **wieder** kühler .
*EN: the new week starts unpredictable and **again** cooler .*
Sign2Text:
DE: **am montag überall** wechselhaft und kühler .
*EN: **on monday everywhere** unpredictable and cooler .*
Sign2Gloss2Text:
DE: die neue woche beginnt wechselhaft und **wechselhaft** .
*EN: the new week starts unpredictable and **unpredictable** .*

Qualitative Results - Alignment Maps

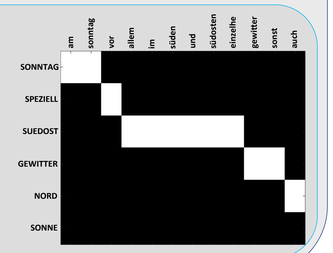
Sign2Text

- Focus on the start of the video.
- Jumps to the end as the sequence finishes.
- Signs = Asynchronous & Different Order



Sign2Gloss2Text

- Cleaner alignment



Conclusions

- Introduced **Phoenix2014T** and **set baselines** for future research.
- **Sign2Text** is a difficult task:
  - **Long term asynchronous dependencies**
  - **spatial embedding** vs. word embedding
- Having **intermediate representations helps.**
- Future work:
  - Learn the intermediate representations in an end-to-end manner.
  - Clever ways to reduce number of frames to shorten sequences.

Acknowledgements

This work was funded by the SNSF Sinergia project "Scalable Multimodal Sign Language Technology for Sign Language Learning and Assessment (SMILE)" grant agreement number CRSII2_160811 and the European Union's Horizon2020 research and innovation programme under grant agreement no. 762021 (Content4All). We would also like to thank NVIDIA Corporation for their GPU grant.

References

[1] X. Chai, G. Li, Y. Lin, Z. Xu, Y. Tang, X. Chen, and M. Zhou, "Sign Language Recognition and Translation with Kinect", IEEE Conference Automatic Face and Gesture Recognition (AFGR), 2013.

[2] J. Bungeroth, and H. Ney, "Statistical Sign Language Translation", Workshop on Representation and Processing of Sign Languages, International Conference on Language Resources and Evaluation (LREC), 2004.

[3] J. Forster, C. Schmidt, O. Koller, M. Bellgardt, and H. Ney. *"Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather"*, International Conference on Language Resources and Evaluation (LREC), 2014.

[4] O. Koller, S. Zargaran, and H. Ney. *"Re-Sign: Re-Aligned End-to-End Sequence Modelling with Deep Recurrent CNN-HMMs"*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.