Universität Rostock Traditio et Innovatio

PLANET artificial intelligence

EUROPÄISCHE UNION
Europäischer Sozialfonds

# Rescoring Sequence-to-Sequence Models for Text Line Recognition with CTC-Prefixes

**Christoph Wick**[1,3], **Jochen Zöllner**[2], **Tobias Grüning**[1]

[1]Planet AI Gmbh Rostock
[2]University of Rostock – Institute of Mathematics
[3] now at Google

# Introduction

## Why combine CTC and S2S Decoding?

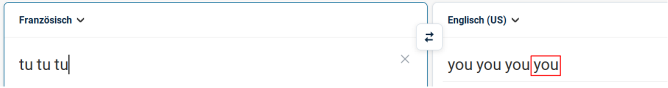
- Sequence to Sequence(S2S) Decoding can perform better due to an intrinsic language model.
- In contrast to Connectionist Temporal Classification(CTC) decoding S2S Decoding has trouble with repetitions.

# Introduction

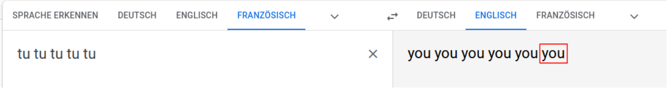
## Why combine CTC and S2S Decoding?

- Sequence to Sequence(S2S) Decoding can perform better due to an intrinsic language model.
- In contrast to Connectionist Temporal Classification(CTC) decoding S2S Decoding has trouble with repetitions.

- Examples from translation models. . .

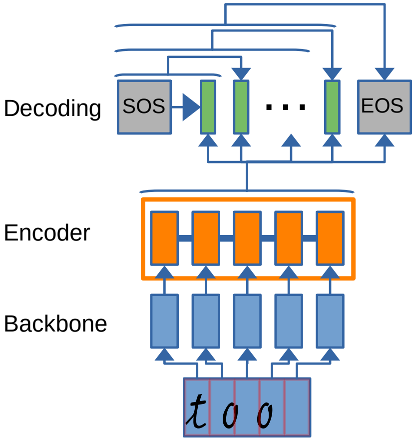



deepl.com

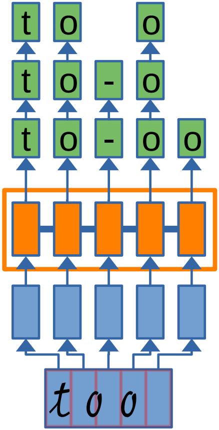



translate.google.com

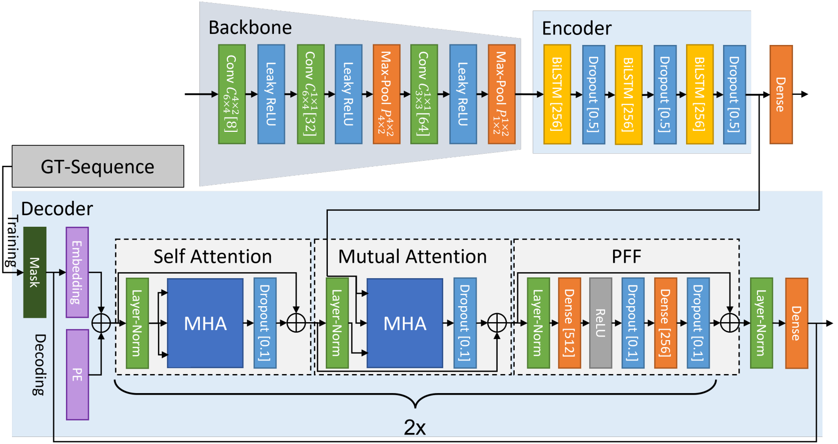
## Sequence to Sequence




## CTC Best Path



# Network-Architecture for Training



# Network-Architecture

- Training loss:

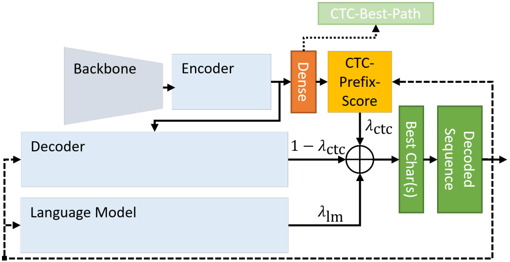
$$L_{\text{tot}} = \lambda_{\text{CTC}} \cdot L_{\text{CTC}} + (1 - \lambda_{\text{CTC}}) \cdot L_{\text{CE}}$$

$L_{\text{CTC}}$: CTC-loss
$L_{\text{CE}}$: Decoder Cross-Entropy-loss
$\lambda_{\text{CTC}} \in [0, 1]$; $\lambda_{\text{CTC}} = 0.3$ for all experiments

# Inference

- Use CTC-Prefix-Score introduced in speech recognition by Watanabe et al. [2017]
- Sequential decoding with beam-search
- With next character language model (traditional transformer)

$$C_{\text{tot}} = \lambda_{\text{CTC}} \cdot C_{\text{CTC}} + (1 - \lambda_{\text{CTC}}) \cdot C_{\text{CE}} + \lambda_{\text{LM}} \cdot C_{\text{LM}}$$



# Datasets

Text line datasets with alphabet size $|A|$ and number of lines in training, validation and test subset

| Dataset | Language | $\|\mathbf{A}\|$ | # Train | # Val | # Test |
|---|---|---|---|---|---|
| IAM | English (en) | 79 | 6,161 | 966 | 2,915 |
| StAZH | Swiss-German (de-ch) | 109 | 12,628 | 1,624 | 1,650 |
| Rimes | French (fr) | 100 | 10,171 | 1,162 | 778 |

# Datasets

Text line datasets with alphabet size $|A|$ and number of lines in training, validation and test subset

| Dataset | Language | $\|\mathbf{A}\|$ | # Train | # Val | # Test |
|---|---|---|---|---|---|
| IAM | English (en) | 79 | 6,161 | 966 | 2,915 |
| StAZH | Swiss-German (de-ch) | 109 | 12,628 | 1,624 | 1,650 |
| Rimes | French (fr) | 100 | 10,171 | 1,162 | 778 |

3 different Next-Character-Language Models(LM) trained on 16 M English, 30 M French and 1 M Swiss-German text lines. Traditional Transformer (character only)

| Dataset | Language | Top-1 | Top-10 |
|---|---|---|---|
| IAM | en | 58.1% | 90.4% |
| StAZH | de-ch | 52.5% | 87.3% |
| Rimes | fr | 58.5% | 92.8% |

# Results: Pretraining with Synthetic Data Only

**CER [%]**

| **Dataset** | **CTC** **Test** | **CTC/Trafo** **Test** |
|---|---|---|
| IAM | 19.5 | 17.3 |
| StAZH | 65.1 | 64.3 |
| Rimes | 25.1 | 23.0 |

- Poor performance without real data.
- CTC/Transformer slightly better than CTC best path decoding

# Results: Influence of Pretrained Models

CTC best path decoding and proposed CTC/Transformer combination

**CER [%]**

| | Pretr. | Test CTC | Test CTC/Tr |
|---|---|---|---|
| IAM | No | 5.47 | 5.10 |
| IAM | Yes | 4.99 | 3.96 |
| StAZH | No | 3.05 | 2.81 |
| StAZH | Yes | 3.06 | 2.66 |
| Rimes | No | 4.31 | 3.88 |
| Rimes | Yes | 4.25 | 3.49 |

- Transformer benefits more from pre-training

# Results: Find Best LM Weigth and Beam Size

$$C_{\text{tot}} = \lambda_{\text{CTC}} \cdot C_{\text{CTC}} + (1 - \lambda_{\text{CTC}}) \cdot C_{\text{CE}} + \lambda_{\textbf{LM}} \cdot C_{\text{LM}}$$

| | **CER [%]** | | | |
|---|---|---|---|---|
| $\boldsymbol{\lambda}_{\textbf{LM}}$ | **0** | **0.1** | **0.5** | **1** |
| IAM | 3.96 | 3.69 | 3.19 | 3.64 |
| StAZH | 2.66 | 2.66 | 2.79 | 3.81 |
| Rimes | 3.49 | 3.40 | 3.39 | 3.57 |

# Results: Find Best LM Weigth and Beam Size

$$C_{\text{tot}} = \lambda_{\text{CTC}} \cdot C_{\text{CTC}} + (1 - \lambda_{\text{CTC}}) \cdot C_{\text{CE}} + \boldsymbol{\lambda_{\text{LM}}} \cdot C_{\text{LM}}$$

| | CER [%] | | | |
|---|---|---|---|---|
| $\boldsymbol{\lambda_{\text{LM}}}$ | **0** | **0.1** | **0.5** | **1** |
| IAM | 3.96 | 3.69 | 3.19 | 3.64 |
| StAZH | 2.66 | 2.66 | 2.79 | 3.81 |
| Rimes | 3.49 | 3.40 | 3.39 | 3.57 |
| **Beams** | **1** | **5** | **10** | **20** |
| IAM | 5.81 | 3.19 | 3.17 | 3.13 |
| StAZH | 4.37 | 2.79 | 2.79 | 2.79 |
| Rimes | 4.10 | 3.39 | 3.19 | 3.19 |

# Ablation and SOTA Comparision on IAM Dataset

| | Authors | Enc. | Dec. | + Data | LM | CER | WER | #P | #/s |
|---|---|---|---|---|---|---|---|---|---|
| A | Ours | LSTM | CTC | No | No | 5.47 | 17.93 | 3.2 | 10.77 |
| B | Ours | LSTM | CTC | Syn | No | 4.99 | 16.85 | 3.2 | 11.57 |
| C | Ours | LSTM | Tr | No | No | 5.61 | 16.24 | 4.8 | 1.90 |
| D | Ours | LSTM | Tr | No | Open | 14.38 | 18.25 | 24 | 0.50 |
| E | Ours | LSTM | Tr | Syn | No | 4.15 | 12.22 | 4.8 | 2.50 |
| F | Ours | LSTM | Tr | Syn | Open | 6.46 | 13.38 | 24 | 0.86 |
| G | Ours | LSTM | CTC/Tr | No | No | 5.09 | 15.88 | 4.8 | 0.69 |
| H | Ours | LSTM | CTC/Tr | No | Open | 4.33 | 12.69 | 24 | 0.37 |
| I | Ours | LSTM | CTC/Tr | Syn | No | 3.96 | 12.20 | 4.8 | 0.70 |
| J | Ours | LSTM | CTC/Tr | Syn | Open | 3.20 | 9.19 | 24 | 0.40 |
| K | Ours (20) | LSTM | CTC/Tr | Syn | Open | **3.13** | **8.94** | 24 | 0.18 |
| L | Ours | LSTM | CTC/Tr | Syn/Val | Open | 3.01 | 8.81 | 24 | 0.42 |
| M | Ours (20) | LSTM | CTC/Tr | Syn/Val | Open | **2.95** | **8.66** | 24 | 0.18 |
| N | Bluche [2] | LSTM | CTC | No | 50K | 3.2 | - | 0.75 | - |
| O | Michael [10] | LSTM | S2S | Val | No | 4.87 | - | - | - |
| P | Yousef [16] | FCN | CTC | No | No | 4.9 | - | 3.4 | - |
| Q | Kang [5] | Tr | Tr | Syn | No | 4.67 | 15.45 | - | - |
| R | Wick [14] | Tr | Bi-Tr | No | No | 5.67 | - | - | - |
| S | Diaz [3] | Tr | CTC | Syn/Real | Open | **2.75** | - | ≈ 12 | - |
| T | Li [7] | Tr | Tr | Syn | No | 3.42 | - | 334 | - |
| U | Li [7] | Tr | Tr | Syn | No | 2.89 | - | 558 | - |

# Ablation and SOTA Comparision on IAM Dataset

| | Authors | Enc. | Dec. | + Data | LM | CER | WER | #P | #/s |
|---|---|---|---|---|---|---|---|---|---|
| A | Ours | LSTM | CTC | No | No | 5.47 | 17.93 | 3.2 | 10.77 |
| B | Ours | LSTM | CTC | Syn | No | 4.99 | 16.85 | 3.2 | 11.57 |
| C | Ours | LSTM | Tr | No | No | 5.61 | 16.24 | 4.8 | 1.90 |
| D | Ours | LSTM | Tr | No | Open | 14.38 | 18.25 | 24 | 0.50 |
| E | Ours | LSTM | Tr | Syn | No | 4.15 | 12.22 | 4.8 | 2.50 |
| F | Ours | LSTM | Tr | Syn | Open | 6.46 | 13.38 | 24 | 0.86 |
| G | Ours | LSTM | CTC/Tr | No | No | 5.09 | 15.88 | 4.8 | 0.69 |
| H | Ours | LSTM | CTC/Tr | No | Open | 4.33 | 12.69 | 24 | 0.37 |
| I | Ours | LSTM | CTC/Tr | Syn | No | 3.96 | 12.20 | 4.8 | 0.70 |
| J | Ours | LSTM | CTC/Tr | Syn | Open | 3.20 | 9.19 | 24 | 0.40 |
| K | Ours (20) | LSTM | CTC/Tr | Syn | Open | **3.13** | **8.94** | 24 | 0.18 |
| L | Ours | LSTM | CTC/Tr | Syn/Val | Open | 3.01 | 8.81 | 24 | 0.42 |
| M | Ours (20) | LSTM | CTC/Tr | Syn/Val | Open | **2.95** | **8.66** | 24 | 0.18 |
| N | Bluche [2] | LSTM | CTC | No | 50K | 3.2 | - | 0.75 | - |
| O | Michael [10] | LSTM | S2S | Val | No | 4.87 | - | - | - |
| P | Yousef [16] | FCN | CTC | No | No | 4.9 | - | 3.4 | - |
| Q | Kang [5] | Tr | Tr | Syn | No | 4.67 | 15.45 | - | - |
| R | Wick [14] | Tr | Bi-Tr | No | No | 5.67 | - | - | - |
| S | Diaz [3] | Tr | CTC | Syn/Real | Open | **2.75** | - | ≈ 12 | - |
| T | Li [7] | Tr | Tr | Syn | No | 3.42 | - | 334 | - |
| U | Li [7] | Tr | Tr | Syn | No | 2.89 | - | 558 | - |

# Ablation and SOTA Comparision on IAM Dataset

| | Authors | Enc. | Dec. | + Data | LM | CER | WER | #P | #/s |
|---|---|---|---|---|---|---|---|---|---|
| A | Ours | LSTM | CTC | No | No | 5.47 | 17.93 | 3.2 | 10.77 |
| B | Ours | LSTM | CTC | Syn | No | 4.99 | 16.85 | 3.2 | 11.57 |
| C | Ours | LSTM | Tr | No | No | 5.61 | 16.24 | 4.8 | 1.90 |
| D | Ours | LSTM | Tr | No | Open | 14.38 | 18.25 | 24 | 0.50 |
| E | Ours | LSTM | Tr | Syn | No | 4.15 | 12.22 | 4.8 | 2.50 |
| F | Ours | LSTM | Tr | Syn | Open | 6.46 | 13.38 | 24 | 0.86 |
| G | Ours | LSTM | CTC/Tr | No | No | 5.09 | 15.88 | 4.8 | 0.69 |
| H | Ours | LSTM | CTC/Tr | No | Open | 4.33 | 12.69 | 24 | 0.37 |
| I | Ours | LSTM | CTC/Tr | Syn | No | 3.96 | 12.20 | 4.8 | 0.70 |
| J | Ours | LSTM | CTC/Tr | Syn | Open | 3.20 | 9.19 | 24 | 0.40 |
| K | Ours (20) | LSTM | CTC/Tr | Syn | Open | **3.13** | **8.94** | 24 | 0.18 |
| L | Ours | LSTM | CTC/Tr | Syn/Val | Open | 3.01 | 8.81 | 24 | 0.42 |
| M | Ours (20) | LSTM | CTC/Tr | Syn/Val | Open | **2.95** | **8.66** | 24 | 0.18 |
| N | Bluche [2] | LSTM | CTC | No | 50K | 3.2 | - | 0.75 | - |
| O | Michael [10] | LSTM | S2S | Val | No | 4.87 | - | - | - |
| P | Yousef [16] | FCN | CTC | No | No | 4.9 | - | 3.4 | - |
| Q | Kang [5] | Tr | Tr | Syn | No | 4.67 | 15.45 | - | - |
| R | Wick [14] | Tr | Bi-Tr | No | No | 5.67 | - | - | - |
| S | Diaz [3] | Tr | CTC | Syn/Real | Open | **2.75** | - | ≈ 12 | - |
| T | Li [7] | Tr | Tr | Syn | No | 3.42 | - | 334 | - |
| U | Li [7] | Tr | Tr | Syn | No | 2.89 | - | 558 | - |

## Summary

- Get benefits of CTC and S2S models
- Competitive error rate with small model
- Eliminating repetition errors
- Slow decoding maybe acceptable depending on use case

## Outlook

- Token-wise decoding for speed up

# References

S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi. Hybrid ctc/attention architecture for end-to-end speech recognition. IEEE Journal of Selected Topics in Signal Processing, 11(8):1240–1253, 2017.

C. Wick, J. Zöllner, and T. Grüning. Rescoring sequence-to-sequence models for text line recognition with ctc-prefixes. In S. Uchida, E. Barney, and V. Eglin, editors, Document Analysis Systems, pages 260–274, Cham, 2022. Springer International Publishing. ISBN 978-3-031-06555-2.