# Sequence Length is a Domain: Length-based Overfitting in Transformer Models

Dušan Variš and Ondřej Bojar

Charles University, Institute of Formal and Applied Linguistics

October 11th, 2021

# Overfitting in Transformers

- Recent models (e.g. GPT-3) increase both in size and in number of training instances.
- We suspect that an overlap in the train-test data could lead to overestimation of model generalization ability.
- Long-range dependencies in transformer:
  - result of poor modeling ability (?) ...
  - ... or lack of data with long-range dependencies?

# Mock Task: String Editing

- Easier evaluation:
  - clear distinction between examples,
  - no ambiguity in correct answers,
  - accuracy metric: exact match with correct solution

| Input | Output |
|---|---|
| push 1 \| 1 0 1 0 | 1 0 1 0 1 |
| reverse – \| 1 0 0 1 1 | 1 1 0 0 1 |

# String Editing: Results

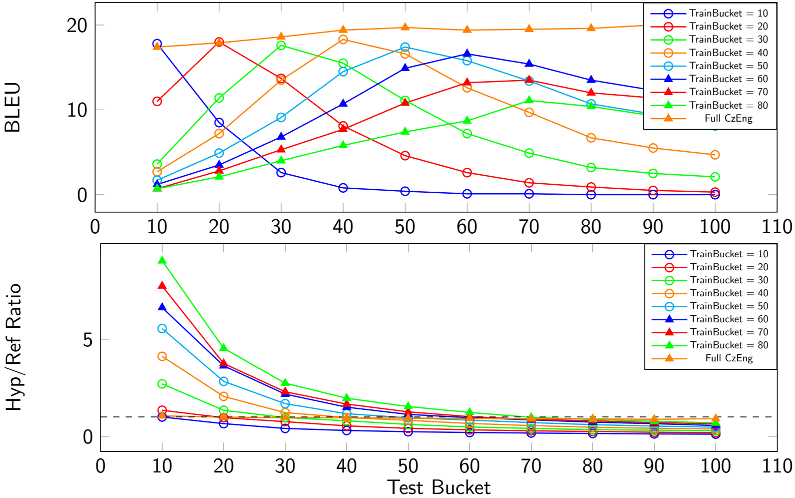
- Training lengths: 10-15
- train/test sequence length mismatch $\rightarrow$ the models fail horribly

| | 0-10 | **11-15** | 16-20 |
|---|---|---|---|
| copy | 62.6 | 100.0 | 0.0 |
| push | 59.1 | 100.0 | 0.0 |
| pop | 0.1 | 100.0 | 0.0 |
| shift | 52.5 | 100.0 | 0.0 |
| unshift | 41.2 | 100.0 | 0.0 |
| reverse | 0.0 | 84.4 | 0.0 |
| all | 15.822 | 97.5 | 0.978 |

# Machine Translation

- Split CzEng 2.0 (Kocmi et al., 2020) into buckets based on target-side (or source-side) sequence length (after subword tokenization).
- Train a separate system on each training bucket.
- Evaluate on WMT newstest split in a similar way.

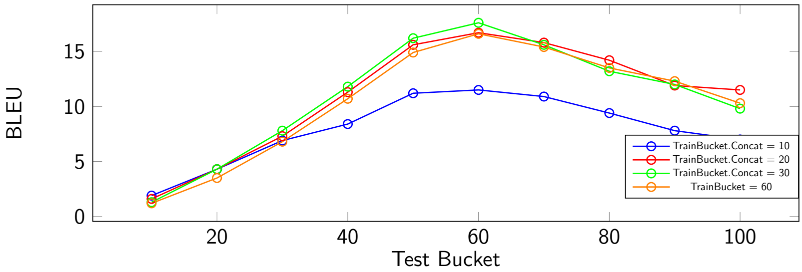
# Machine Translation: Results (Target-length Buckets)



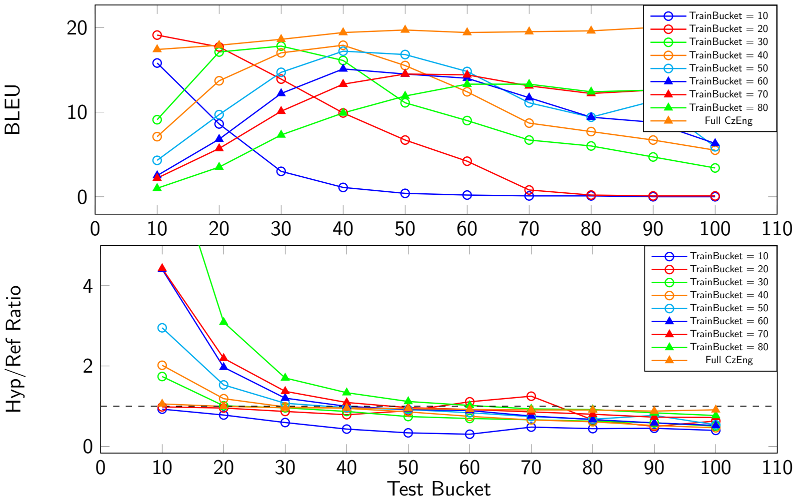
# Machine Translation: Synthetic concatenated data

- Create a synthetic 60-bucket data using concatenation of:
  - 6 x 10-bucket sentences,
  - 3 x 20-bucket sentences,
  - 2 x 30-bucket sentences.
- We concatenate consecutive sentence pairs (after shuffling).
- Compare with the system trained on 60-bucket data.

# Synthetic Concatenation: Results (Target-length Buckets)



# Machine Translation: Source-length Buckets



# Source-length Buckets: Target-length Distributions

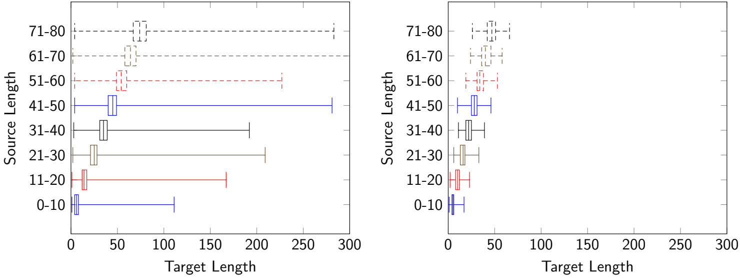



Figure: **Left:** Train Distribution, **Right:** Test Distribution

# See You at the Posters!

## Sequence Length is a Domain: Length-based Overfitting in Transformer Models

Dušan Variš and Ondřej Bojar
varis,bojar@ufal.mff.cuni.cz

Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic

### Introduction

Transformers generalize poorly to longer AND shorter sequence editing examples. Similar trends can be observed on MT task.

### Methods

Split data to buckets based on target-side length. Train a separate NMT system on each training bucket and evaluate it on the validation buckets.

### Results

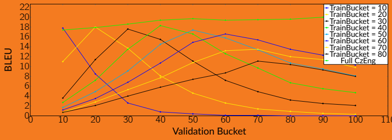
Strong implication of target-side-length overfitting in Transformers that use absolute position encoding.

- Higher train-test length difference → higher performance drop.
- Hypothesis length similar to that of training data.
- Length overfitting could be avoided with relative position embeddings (Neishi and Yoshinaga, 2019).

Transformers with **absolute position encoding** output sequences of **length similar** to sequences in **training data.**

| | |
|---|---|
| Source (30-bucket) | The company does not collect its mail and it has closed its official headquarters in Žižkov more than six years ago. |
| Hyp1 (10-bucket) | Společnost nesbírá poštu a zavřel oficiální sídlo. |
| *Hyp1 (gloss)* | *The company does not gather mail and closed official headquarters.* |
| Hyp2 (30-bucket) | Společnost neshromažďuje poštu a již před více než šesti lety zavřela své oficiální sídlo v Žižkově. |
| *Hyp2 (gloss)* | *The company does not collect mail and more than six years ago closed its official headquarters in Žižkov.* |
| Hyp3 (60-bucket) | Společnost nevybírá poštu a uzavřela své oficiální sídlo v Žižkově více než šest let ago. v Žižkově. Společnost neshromažďuje poštu a uzavírá oficiální ústředí v Žižkově více než šest let agr. o. |
| *Hyp3 (gloss)* | *The company does not pick up mail and closed up its official its official headquarters in Žižkov more than six years ago. in Žižkov. The company does not collect mail and closes up official headquarters in Žižkov more than six years agr. o.* |
| Reference (30-bucket) | Nepřebírá poštu a oficiální sídlo na Žižkově zrušila před více než šesti lety. |
| *Ref (gloss)* | *(The company) does not collect mail and official headquarters in Žižkov closed up more than six years ago.* |

**Take a picture to download the full paper**




This work was supported by the GA ČR NEUREM3 grant (Neural Representations in Multi-modal and Multi-lingual Modelling, 19-26934X (RIV: GX19-26934X)), by the SSHOC grant (Social Sciences & Humanities Open Cloud, 823782) and by SVV 260 453 grant. This template is based on Mike Morrison's idea of Better Poster. It was modified to latex by Rafael Bailo and modified for UFAL purposes by Tom Kocmi.