A Great Toolkit is Just the Beginning:

Learnings From Building Amazon-Scale Production NMT Systems

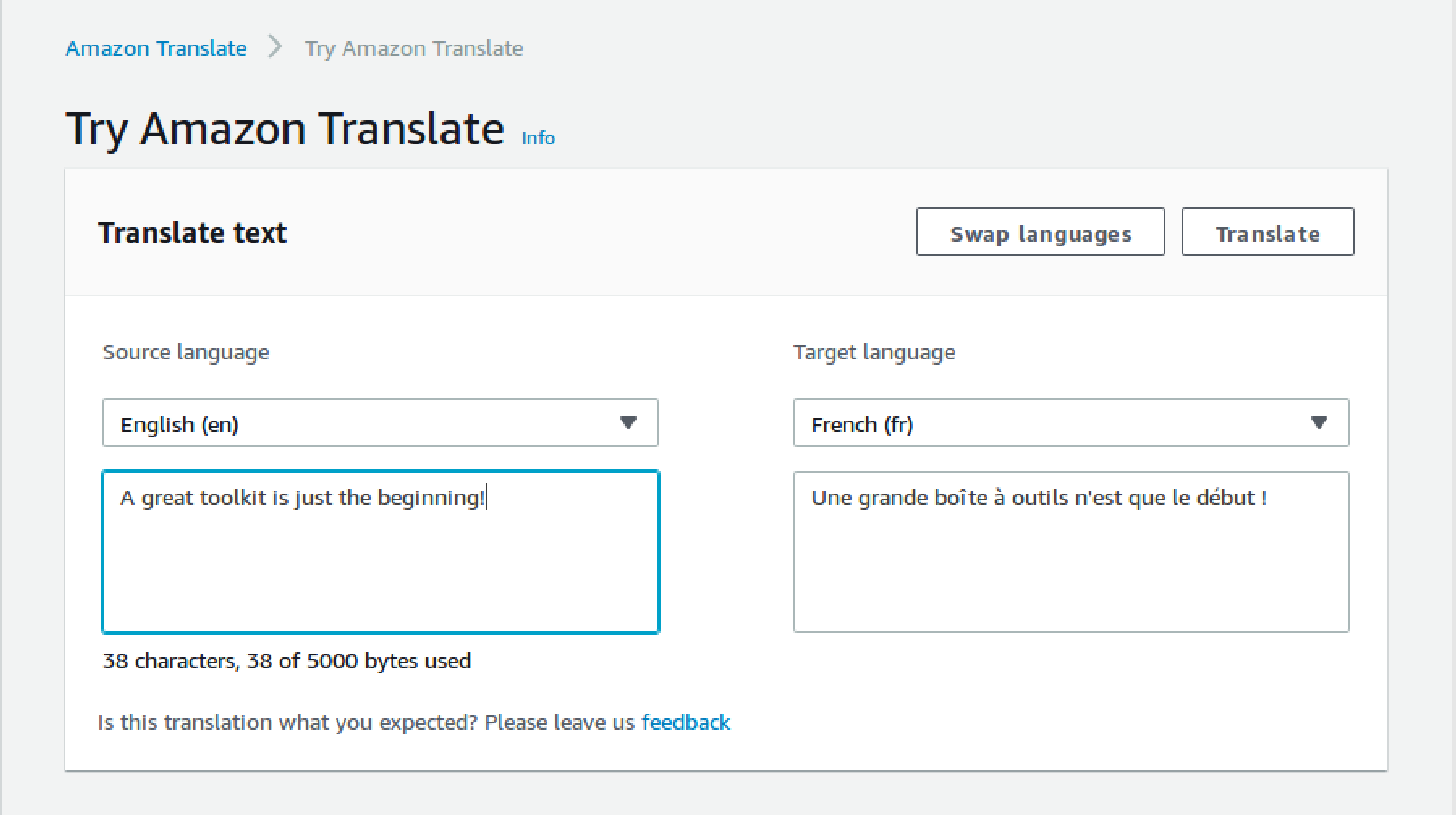
Greg Hanneman, Ann Clifton, Silja Hildebrand, Patrick Porter, Steve Sloto
Scarlett MT Research and System-Building Teams
Translation Services and Products

Amazon Machine Learning Conference
April 26, 2018

Talk Overview

- **Background**
  - Amazon Translate
  - NMT with Sockeye
  - Academic research community
- Differences of configuration
- Constraints of production
- Summary

Amazon Translate



Amazon Translate

- General-purpose MT: EN ↔ AR, DE, ES, FR, PT, ZH
- Started April 2017; preview Nov. 2017; GA April 2018
- System building: ≈400 experiments in six months
- Tons of moving parts: Core ML, TSP, InTech, AWS AI
- AMLC paper: “Amazon Translate: A Cross-Organization Collaborative Success Story”

NMT with Sockeye

- Amazon’s open-source NMT toolkit
- Core “decoder” in Amazon Translate systems
- Featureful, production-ready, state-of-the-art
  - Implements three NN architectures for NMT
  - Based on MXNet
  - Extremely configurable...

NMT with Sockeye

origin_train_src
origin_train_tgt
origin_adapt_src
origin_adapt_tgt
origin_dev_src
origin_dev_tgt
origin_dev_adapt_src
origin_dev_adapt_tgt
origin_test_src
origin_test_tgt
nmt_encoder
nmt_decoder
nmt_wordrep_size_src
nmt_wordrep_size_tgt
rnn_encoder_layers
rnn_decoder_layers
rnn_layer_size
rnn_cell_type
rnn_layer_normalization
rnn_residual_connections
rnn_attention_type
rnn_attention_size
rnn_attention_use_prev_word
rnn_attention_feed_context
rnn_attention_cov_type
rnn_attention_cov_num_hidden
rnn_decoder_state_init

rnn_encoder_reverse_input
transformer_layers
transformer_model_size
transformer_attention_heads
transformer_feed_fwd_num_hidden
transformer_preprocess
transformer_postprocess
transformer_pos_emb_typecnn_layers
cnn_num_hidden
cnn_kernel_width
cnn_hidden_dropout
cnn_activation_type
cnn_positional_embedding_type
cnn_weight_normalization
vocab_weight_tying
vocab_word_min_count_src
vocab_word_min_count_tgt
vocab_words_src
vocab_words_tgt
train_embed_dropout
train_rnn_dropout_inputs
train_rnn_dropout_states
train_rnn_dropout_recurrent
train_rnn_decoder_hidden_dropout
train_transformer_dropout_attn
train_transformer_dropout_relu
train_transformer_dropout_prepost

adapt_embed_dropout
adapt_rnn_dropout_inputs
train_min_len
train_max_len_src
train_max_len_tgt
train_bucket_width
train_additional_args
train_optimizer
train_loss
train_sce_alpha
train_normalize_loss
train_clip_gradient
train_optimized_metric
train_num_monitor_bleu
train_metric_max
train_batch_type
train_batch_size
train_fill_up
train_learning_rate
train_checkpoint_frequency
adapt_checkpoint_frequency
train_rate_schedule
train_rate_decay_when
train_rate_decay
train_rate_warmup_steps
train_early_stop_when
train_min_num_epochs

train_device_ids
train_average_strategy
train_average_num_checkpoints
train_ensemble_size
decode_beam_size
decode_length_penalty_alpha
decode_length_penalty_beta
decode_max_input_len
decode_device_id
decode_max_output_len_num_stds
decode_bucket_width_source
decode_bucket_width_target
decode_additional_args
spm_vocab_size
spm_vocab_type
spm_model_type
spm_normalization_rule_name
spm_user_defined_symbols
bpe_num_operations
bpe_vocab_threshold

Academic Research Community

- Yearly shared MT task, other publications
- Compared to Amazon Translate…
  - Data is quite small
  - Domains are quite limited
  - Timing is very generous
  - Risk after failure is low

Academic Research Community

- Yearly shared MT task, other publications
- Compared to Amazon Translate…
  - Data is quite small
  - Domains are quite limited
  - Timing is very generous
  - Risk after failure is low

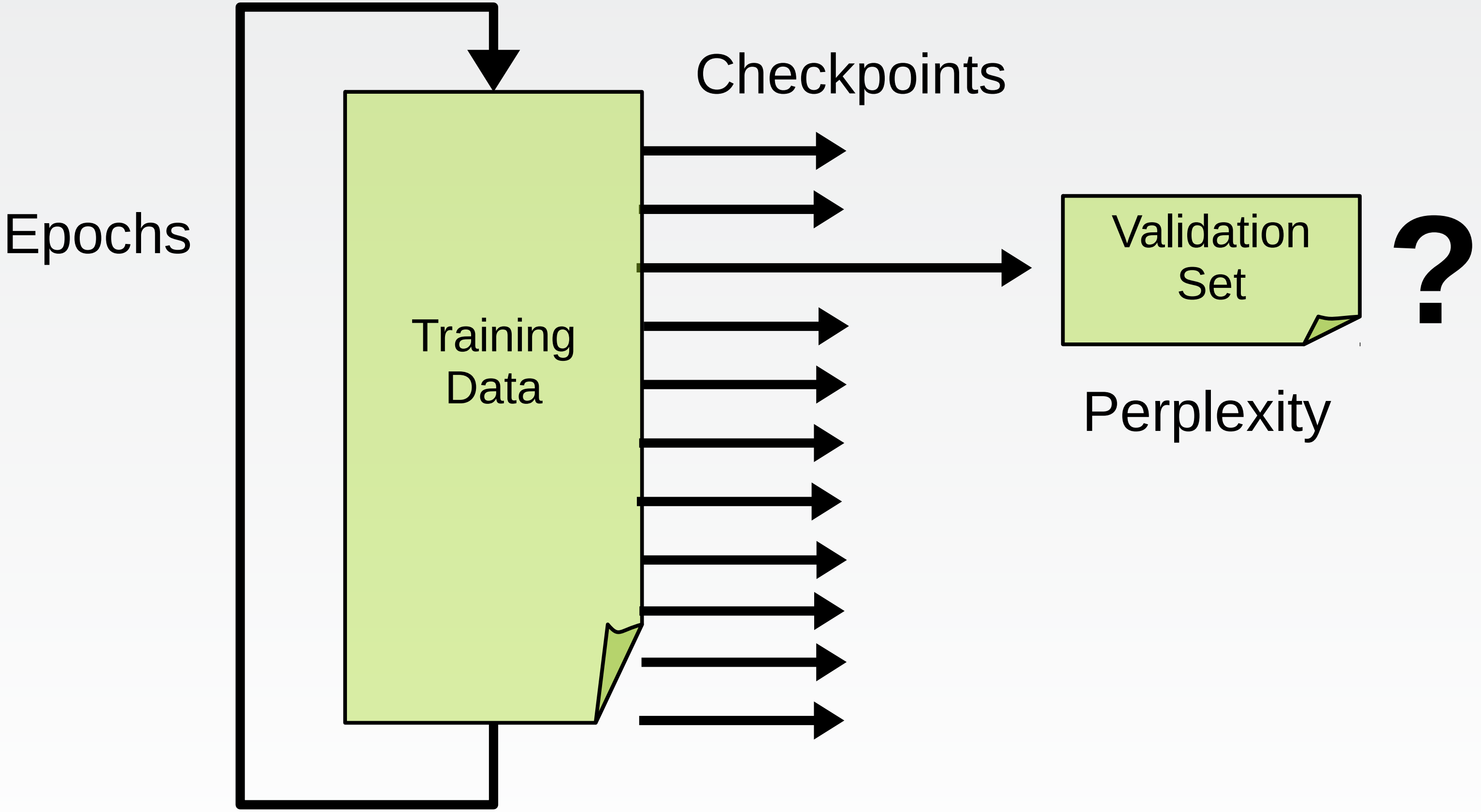
Differences of configuration

Constraints of production

Talk Overview

- Background
- **Differences of configuration**
  - Training stopping criterion
  - Learning rate decay
- Constraints of production
- Summary

NMT Training Procedure

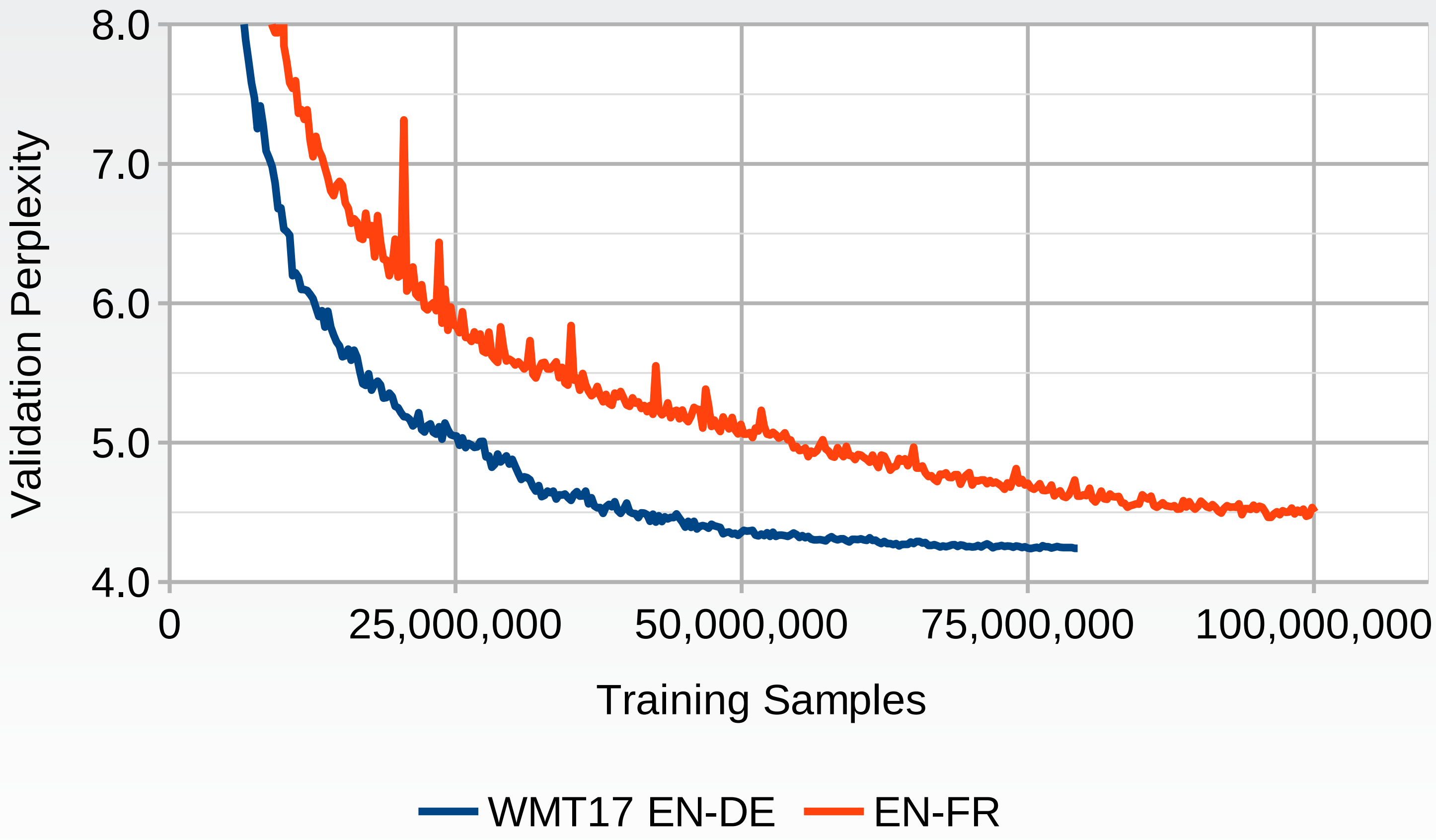


Training Stopping Criterion

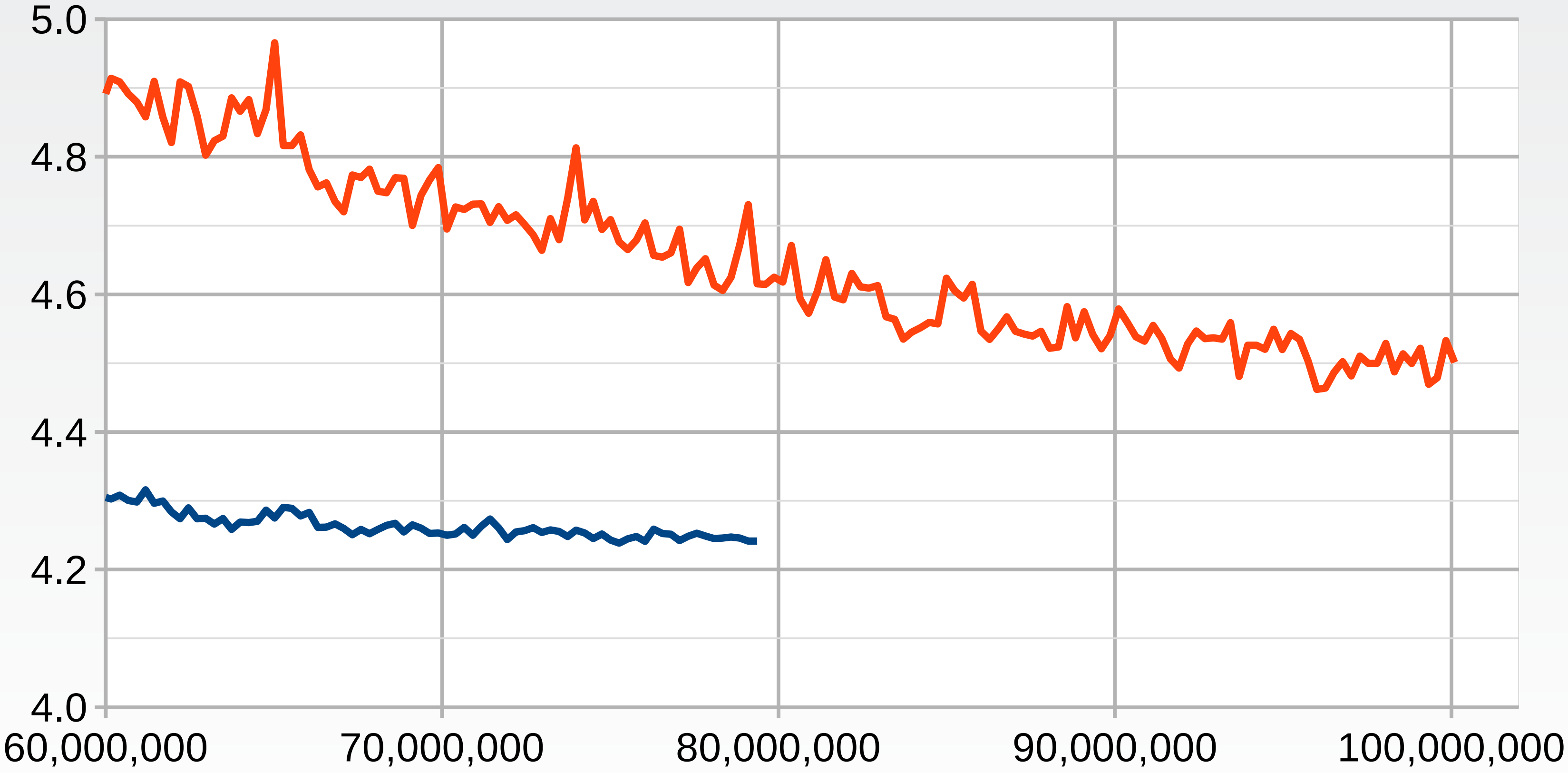
- Two main options:
  - Perplexity stops improving on validation set
  - Fixed number of samples on training data

- Does it depend on the size of training data?

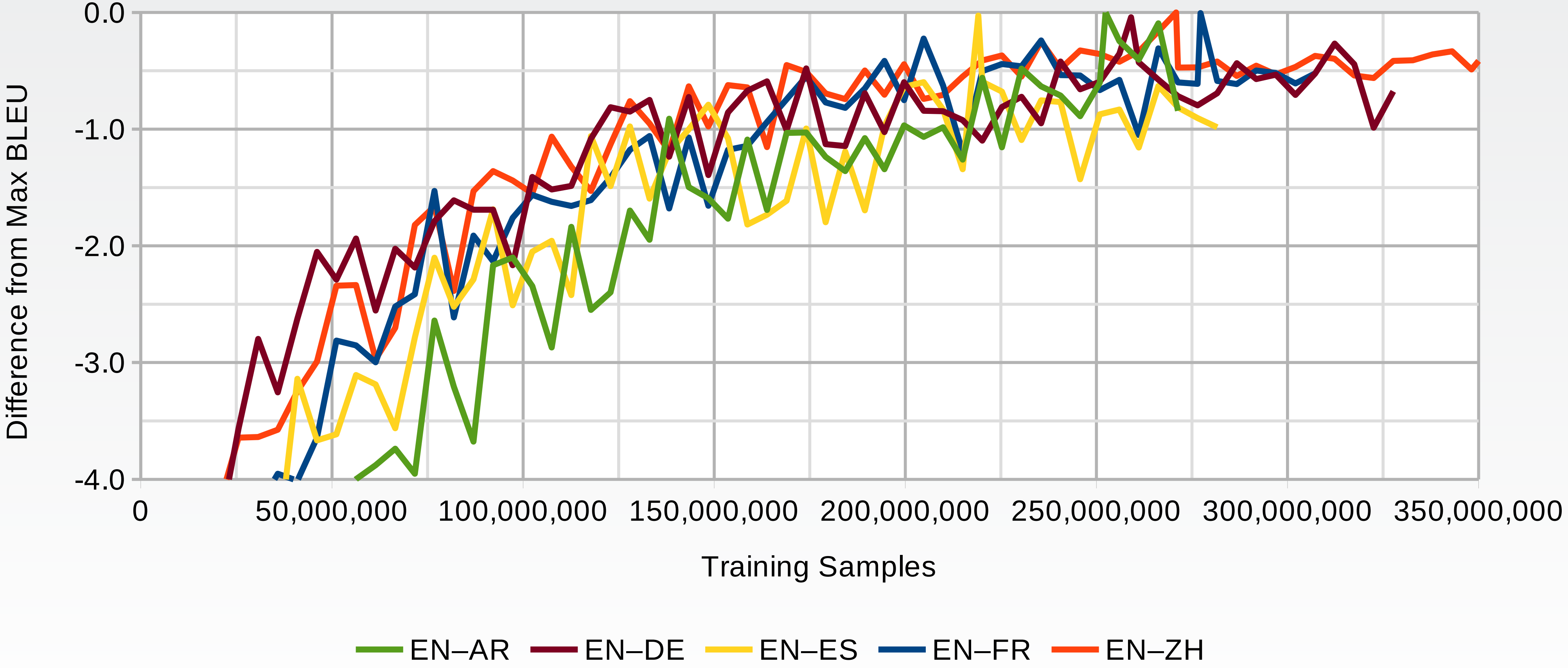
Training Stopping Criterion



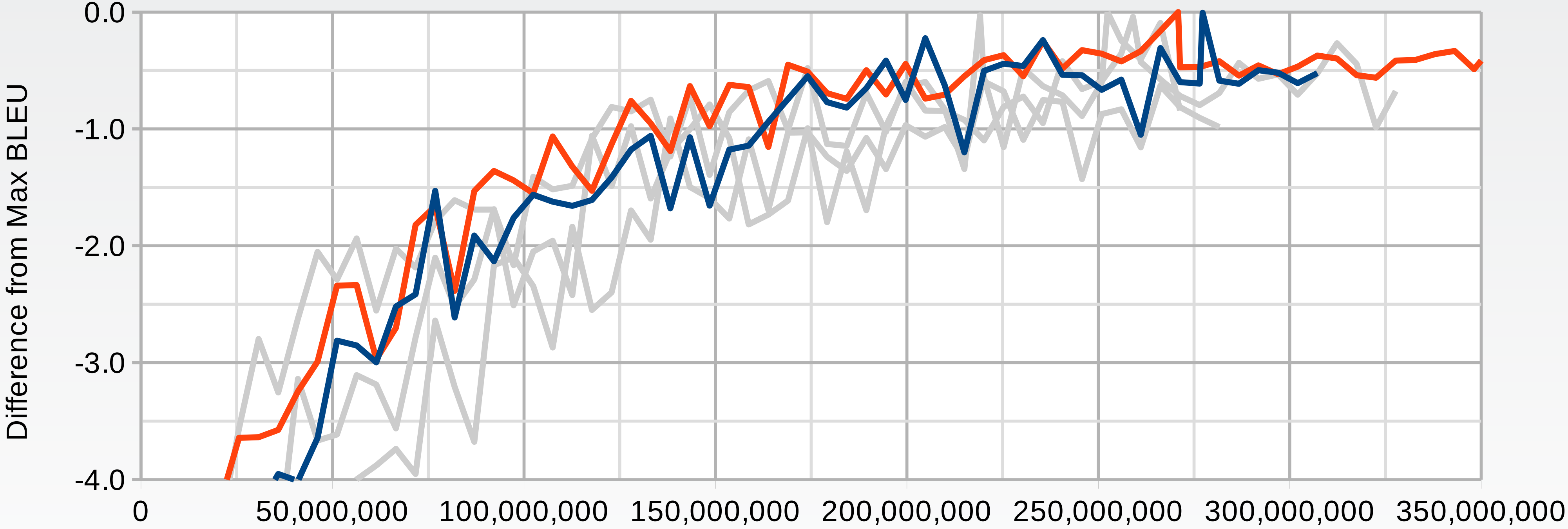
Training Stopping Criterion



Training Stopping Criterion



Training Stopping Criterion

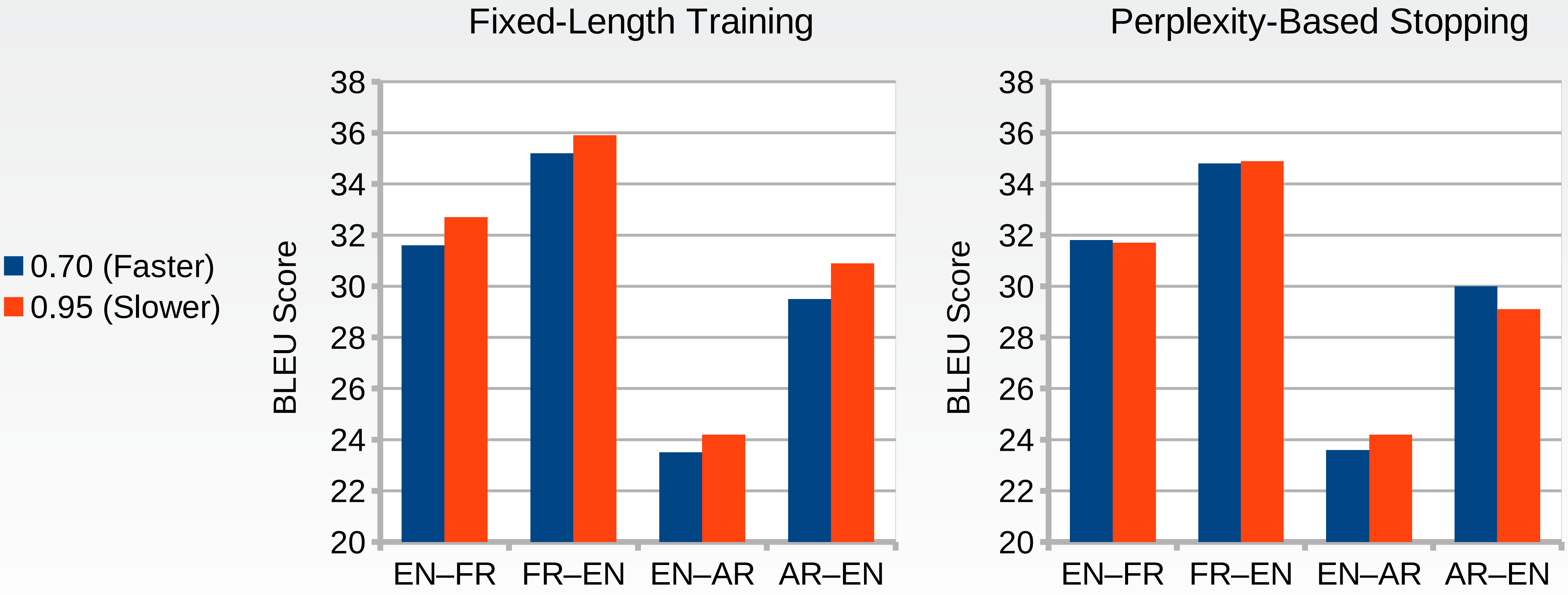


French has 2.5 times as much data as Chinese

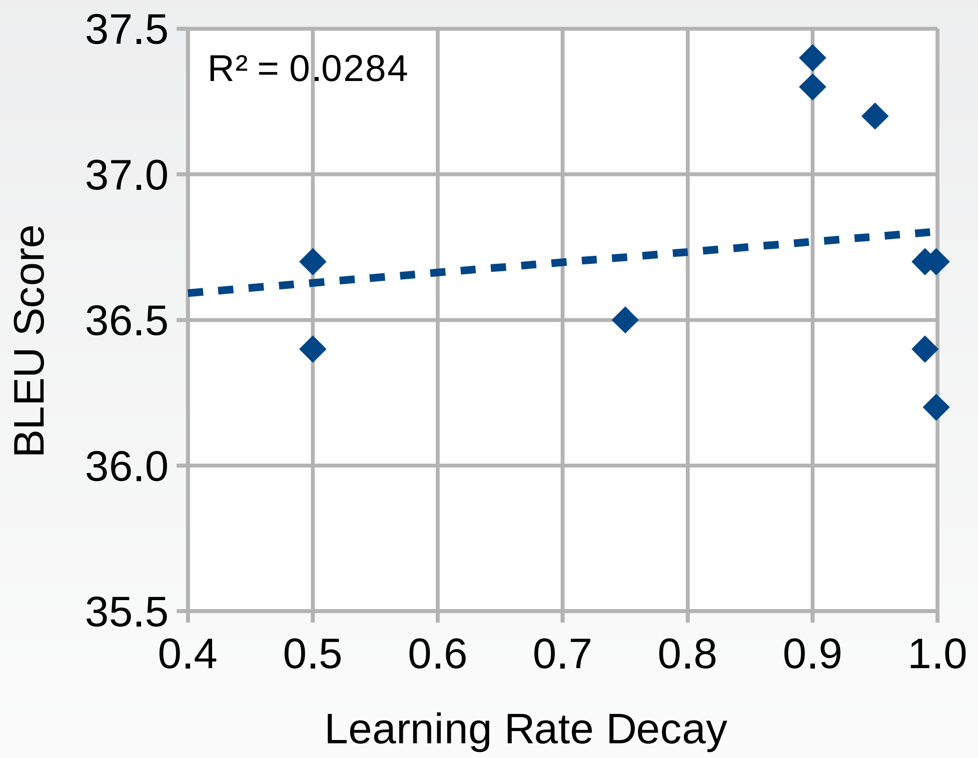
Learning Rate Decay

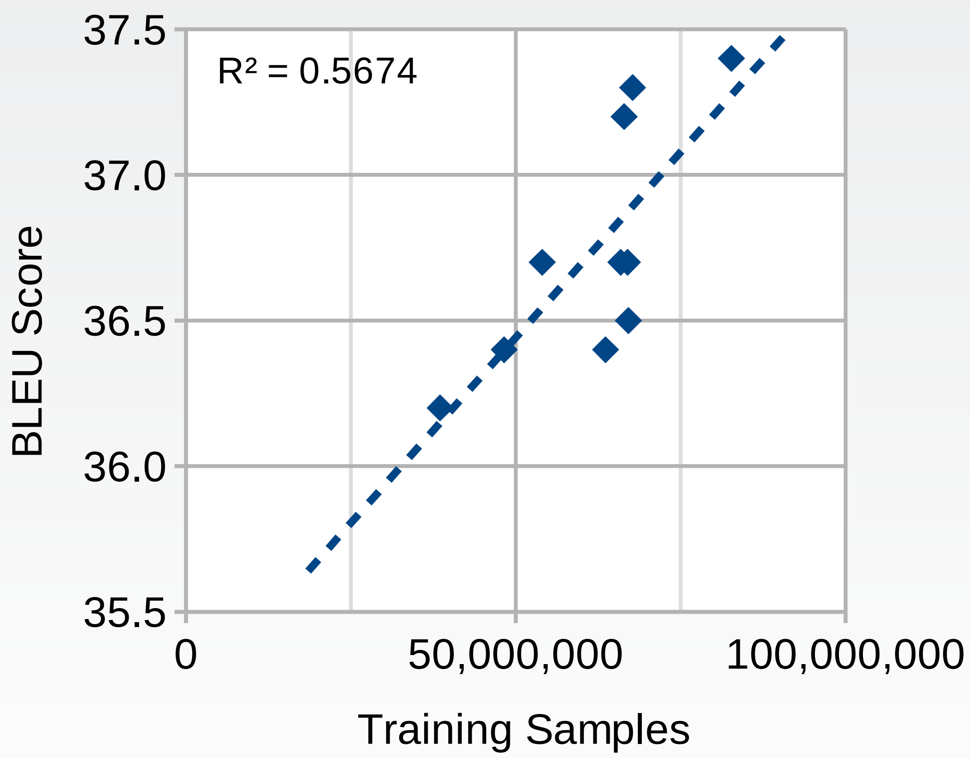
- Too quickly:
  - Get trapped before seeing enough data
  - Premature perplexity-based trigger
- Too slowly:
  - Can’t probe areas in fine detail

Learning Rate Decay



Learning Rate Decay





Learnings: Configuration Differences

- On our data, samples matter more than epochs
- Variable-length training is an experimental confound

  We train all systems to a fixed number of samples
- Slow learning rate decay performs better
- Fast decay doesn't hurt convergence time

  We use slow decay

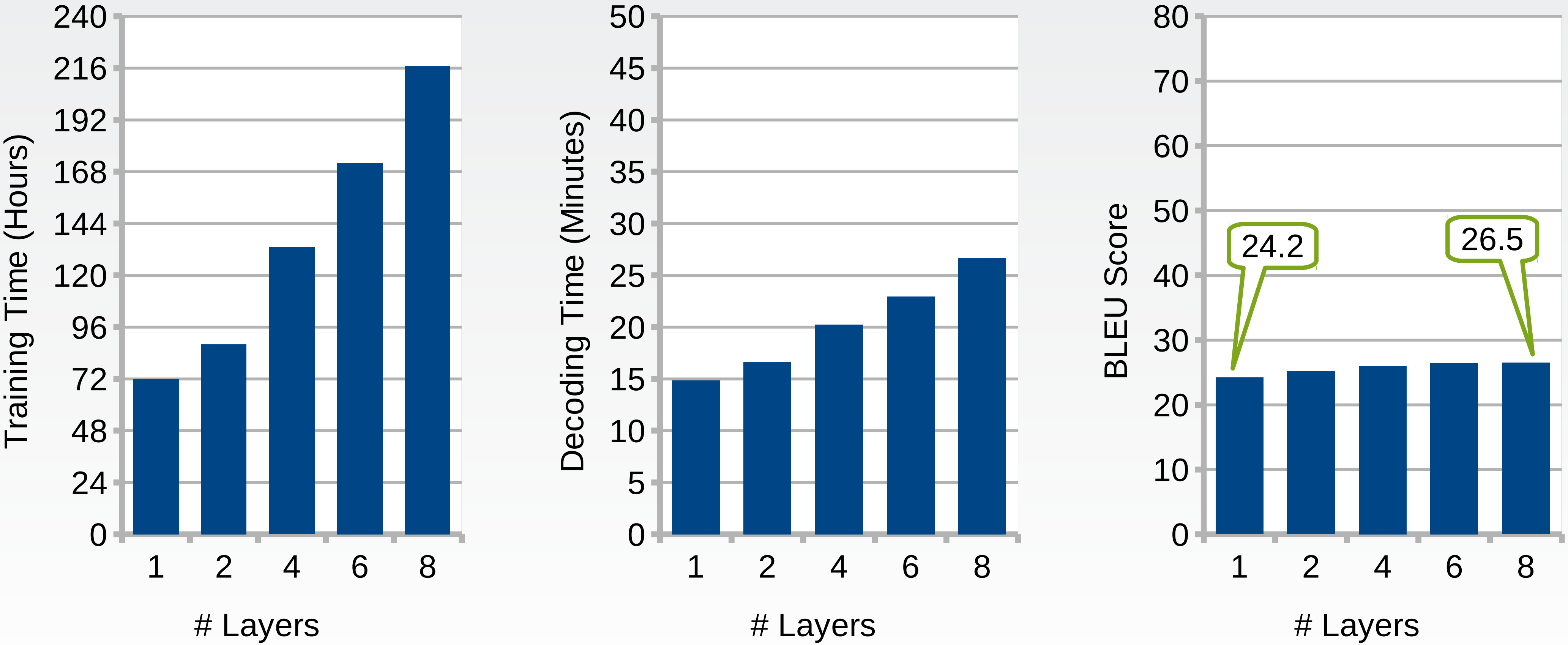
Talk Overview

- Background
- Differences of configuration
- **Constraints of production**
  - Number of RNN layers
  - Embarrassing failures
- Summary

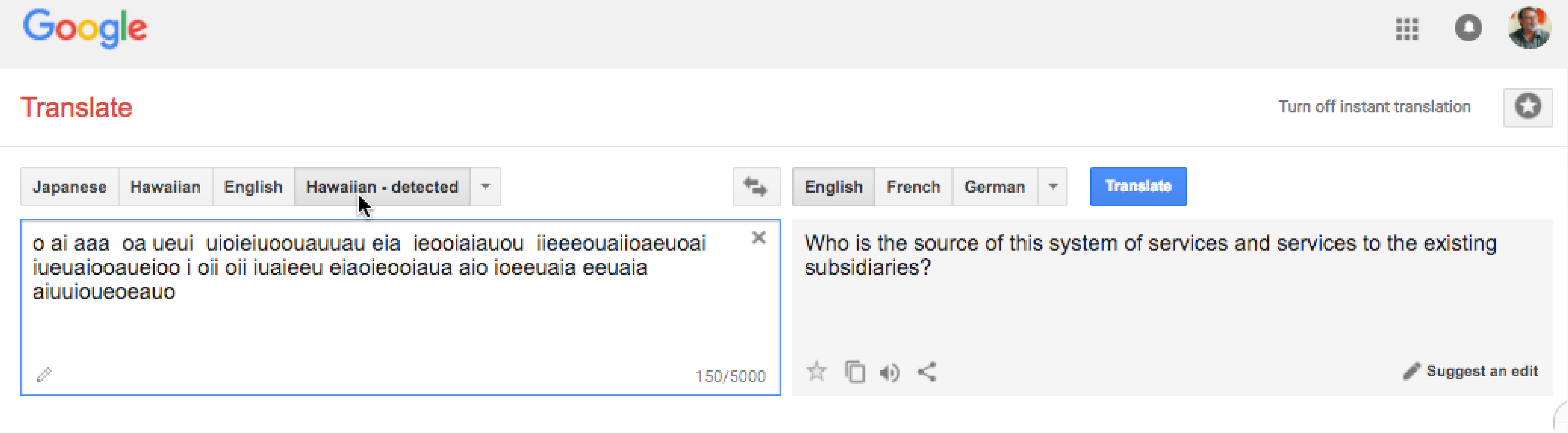
Number of RNN Layers

- Shallow models train and decode faster
- Deep models produce better output

Number of RNN Layers



Embarrassing Failures



Language Log, Feb. 17, 2018

Embarrassing Failures

aporellos	⟶	seaaaaaaaaaaaaaaaaa
emociónnecesaria	⟶	emocionyyyyy
MNN20FCSI7	⟶	MNNN20FCSI7
alsuyo	⟶	allohhhhhhhhhhhhhhh
fuequeprimero	⟶	Flirrrrantfirst
informático	⟶	IT IT
Designer	⟶	Designer Designer Designer Designer Designer
12,35	⟶	12,35 12,35
parquepara	⟶	para para para
ELLA	⟶	LA LA LA LA LA LA LA LA LA LA LA LA LA LA LA LA

Embarrassing Failures

aporellos → seaaaaaaaaaaaaaaaaa
emociónnecesaria → emocionyyyyy
MNN20FCSI7 → MNNN20FCSI7
alsuyo → allohhhhhhhhhhhhhh
fuequeprimero → Flirrrrantfirst

Three or more repeated characters that aren't in the input

informatico → IT IT
Designer → Designer Designer Designer Designer Designer
12,35 → 12,35 12,35
parquepara → para para para
ELLA → LA LA LA LA LA LA LA LA LA LA LA LA LA LA LA LA

Single-word input

Repeated tokens

Embarrassing Failures

- Due to scarcity of single-word training examples?
- More likely for unknown words?
- More likely for words translated in small chunks?
- More likely for words never seen as sentence-final?
- …?

Embarrassing Failures

- ✔ Due to scarcity of single-word training examples?
- ✘ More likely for unknown words?
- ✘ More likely for words translated in small chunks?
- ✔ More likely for words never seen as sentence-final?
- ✔ Most likely for non-translatable placeholders

Learnings: Production Constraints

- More compute power isn't always a real-time win

  We use two-layer RNNs for best quality × speed

- Model unexpectedly good at unknown/rare words; unexpectedly bad at frequent words with no context

  We added more single-word examples to training

  We tried a more nuanced pre-processing approach with a targeted evaluation

Talk Overview

- Background
- Differences of configuration
- Constraints of production
- **Summary**

Summary

- Learnings relevant to other ML domains
  - Data size vs. training length vs. experimental confounds
  - Speed vs. quality trade-offs
  - Validate assumptions: analysis can yield surprising results

Thank you!

Thanks also to:

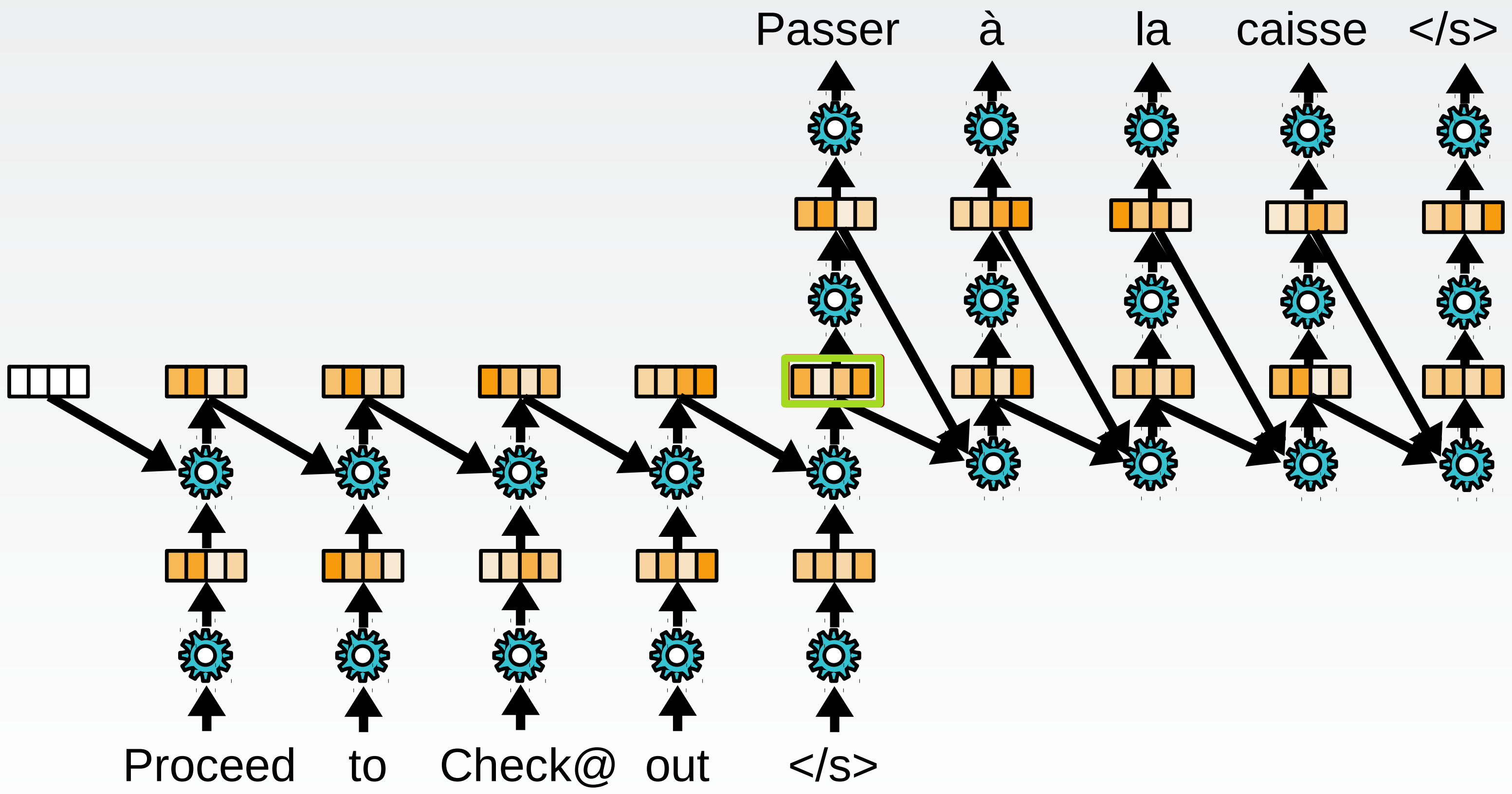
- Chris Jordan-Squire (ex-Amazon; experiments)
- Alon Lavie (feedback and guidance)
- Pittsburgh/Berlin MT Research team (Sockeye)

Term Masking

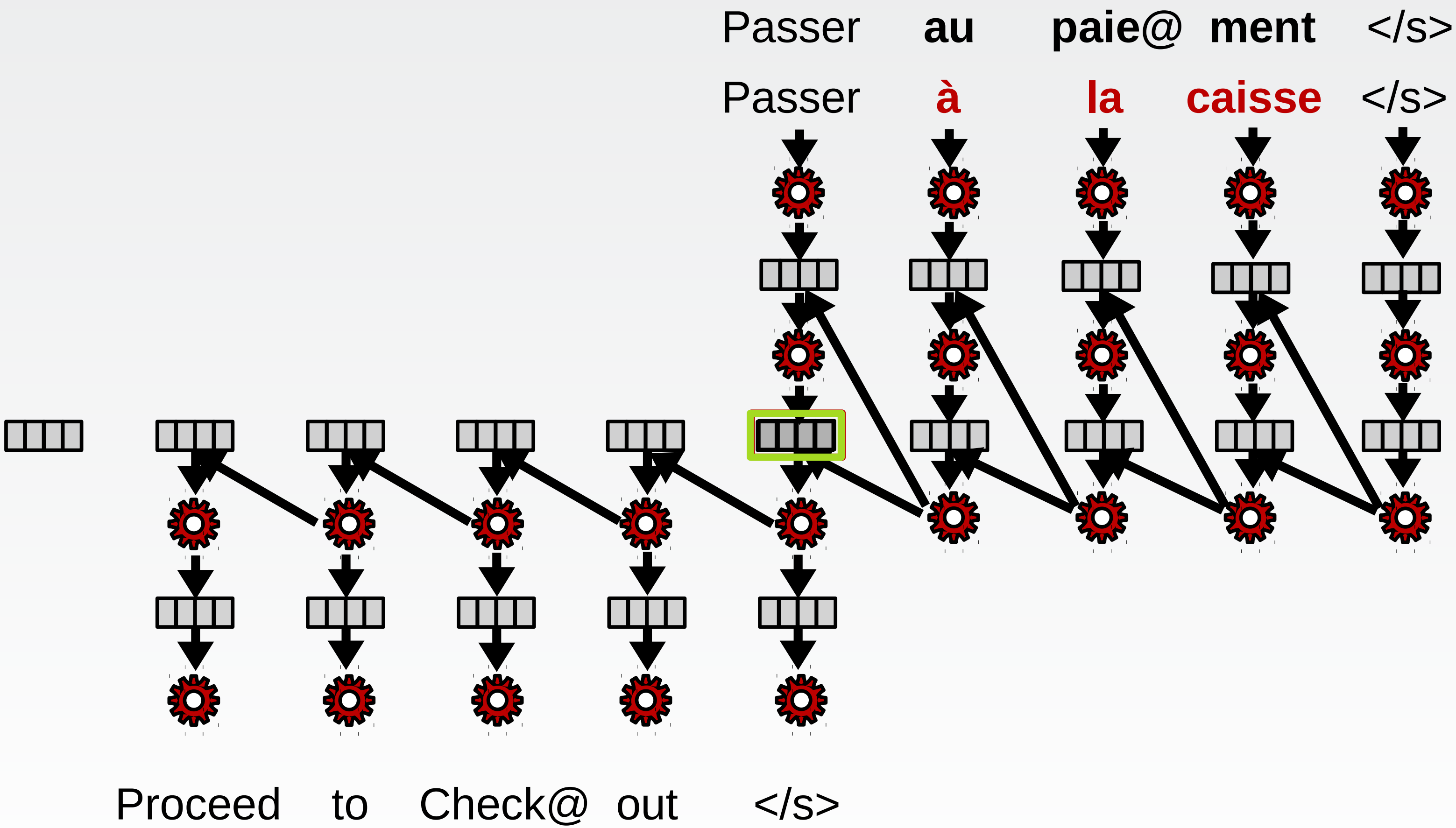
- Anonymize certain non-translatable tokens

```
Idea #317: @MrFixIt said to try www.isitdown.com.

Idea #NUM: HANDLE said to try URL.

Idée No. NUM: HANDLE a dit d'essayer URL.

Idée No. 317: @MrFixIt a dit d'essayer www.isitdown.com.
```

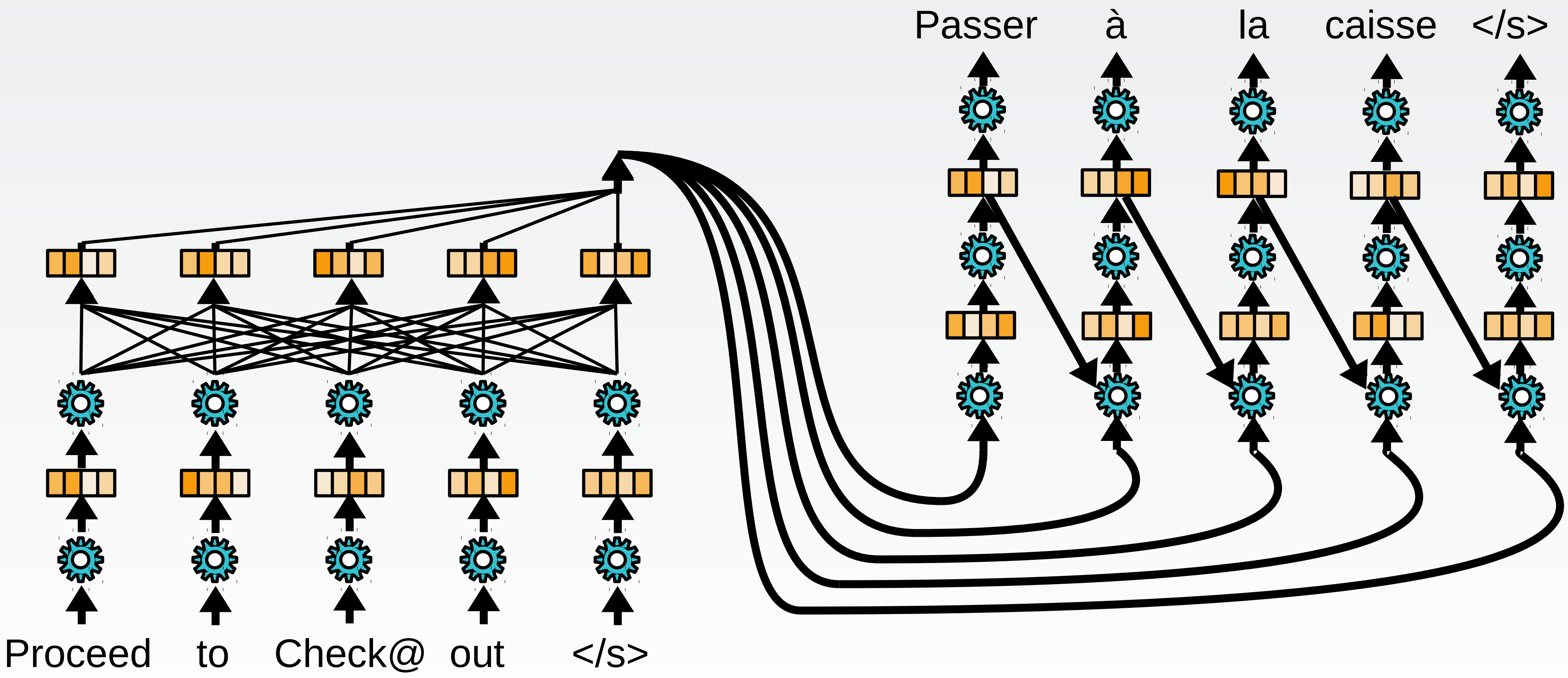
RNN Training for NMT



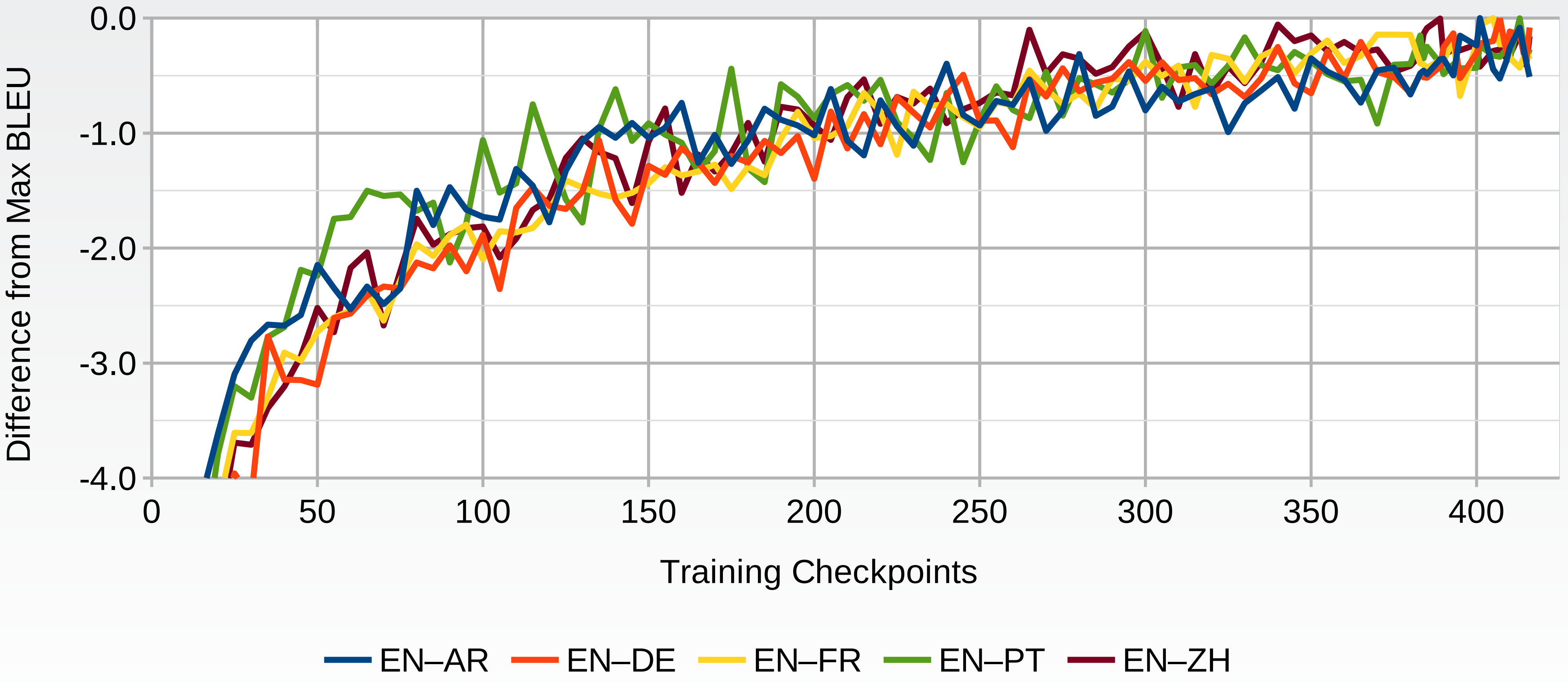
RNN Training for NMT



Transformer Training for MT



Training Convergence: Transformer



Amazon Translate Quality