Article Review : 2024-Fall | ASY6001 | 딥러닝

Transforming Machine Translation:

A Deep Learning System Reaches News Translation Quality Comparable to Human Professionals

Martin Popel, Marketa Tomkova, Jakub Tomek et al.
(Charles University, University of Oxford, Google Brain)

Nature Communications 11, 4381 (2020)

2024266791 컴퓨터·소프트웨어학과 윤 예 진

Transforming Machine Translation

Contents

01
Background; Pre-requisite

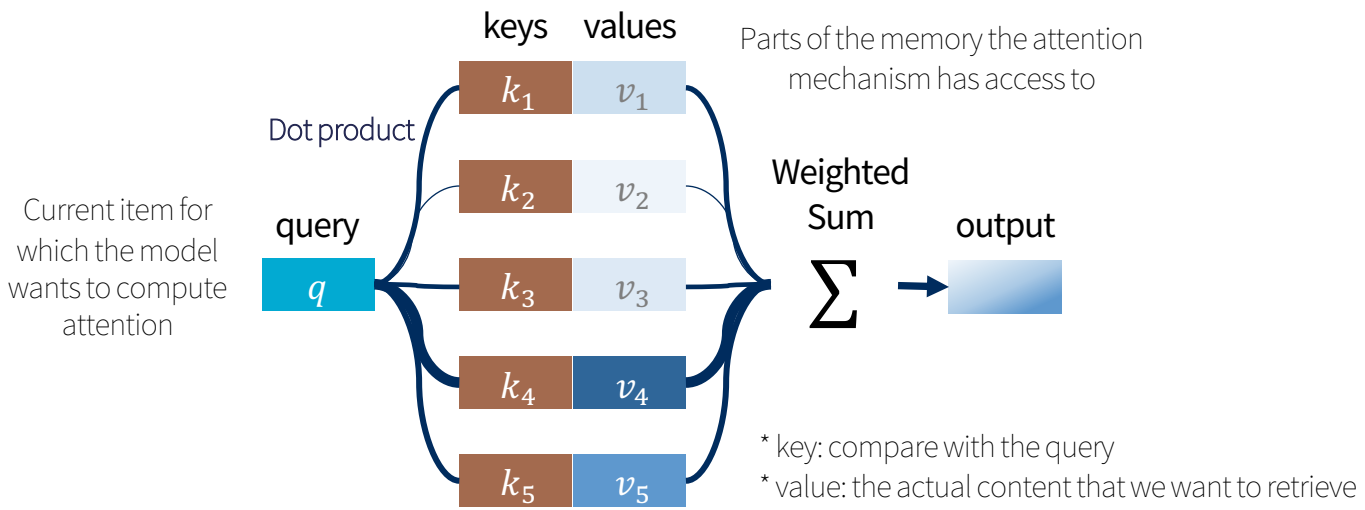
(Self-)Attention Mechanisms

Transformers

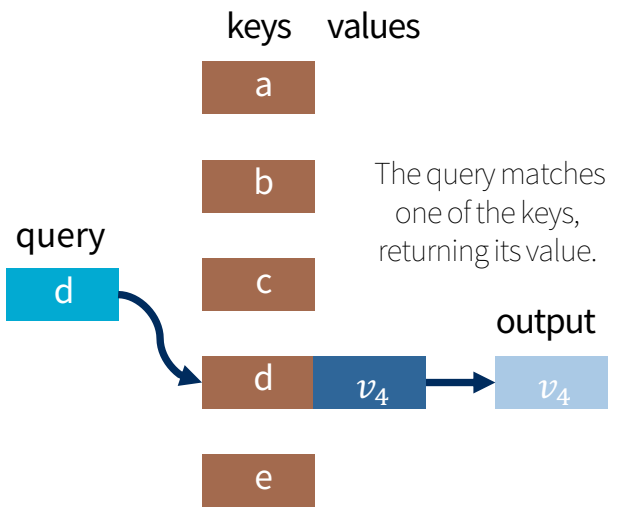
Multi-head Attention

The Concept of Attention Mechanisms

- **Attention is just a weighted average**
  - Weighted Average
    - Query Matching: The query matches all keys 'softly', to a weight between 0 and 1.
    - Weighted Sum: The keys' values are multiplied by the weights and summed.



[Weighted Average]

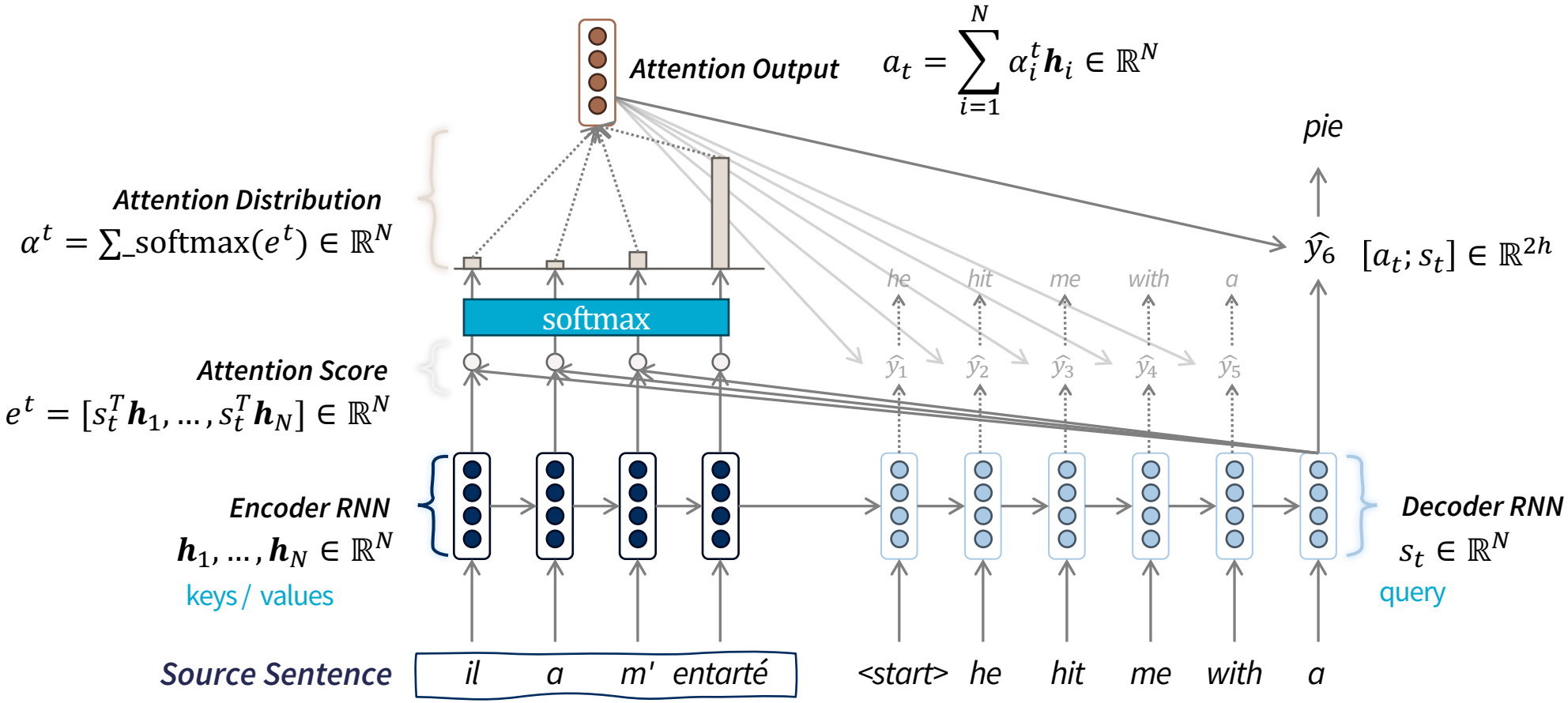


[Lookup Table]

On each step of the decoder, use *direct connection to the encoder* to focus on a particular part of the source sequence

Sequence-to-sequence with Attention

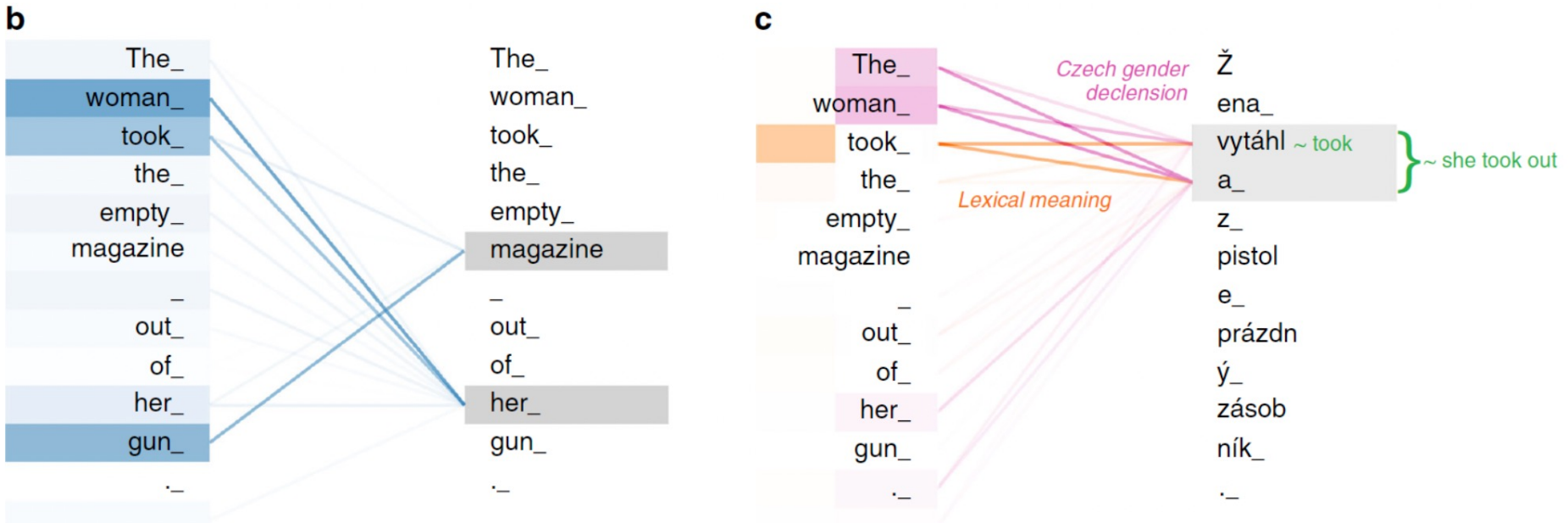
- **Example of Neural Machine Translation**



On each step of the decoder, use *direct connection to the encoder* to focus on a particular part of the source sequence

The Concept of Self-Attention Mechanisms

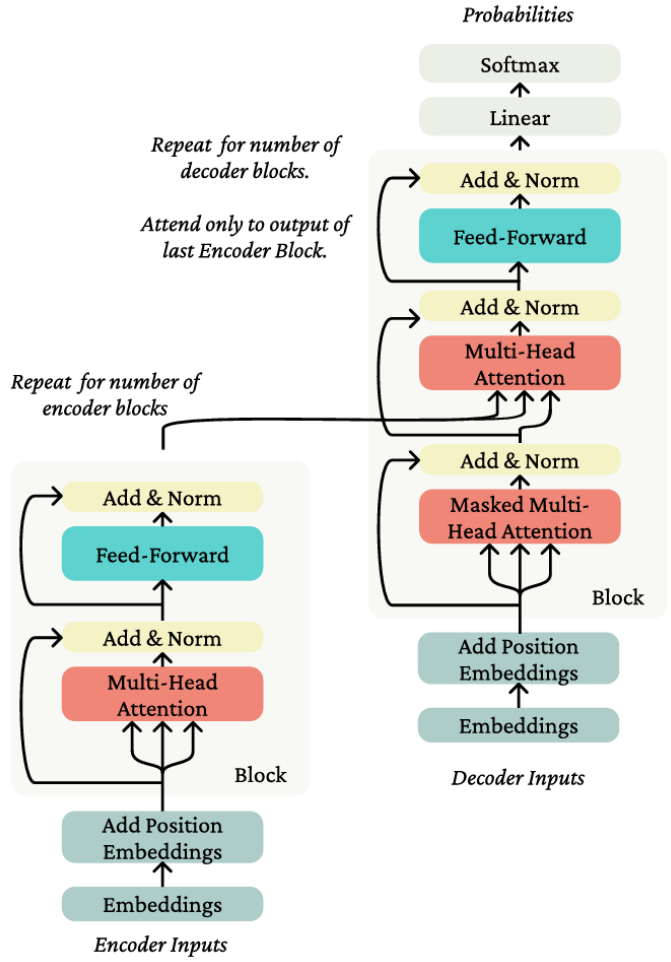
- **Visualization of Self-attention between the first 2 layers**
  - One attention head shown, focusing on “magazine” & “her”
    - Strong attention link between ‘magazine’ and ‘gun’ → zásobník (Gun magazine), not časopis (news magazine)
  - Another attention link between ‘women’ and ‘her’
    - how the system internally learns coreference



On each step of the decoder, use *direct connection to the encoder* to focus on a particular part of the source sequence

Wrap-up: the Block of Transformers

- **Self-attention**: the basis of the method
- **Position representations**: Specify the sequence order, since self-attention is an unordered function of its inputs
- **Nonlinearities**: at the output of the self-attention block, frequently implemented as a simple feed-forward network
- **Masking**: to parallelize operations while not looking at the future



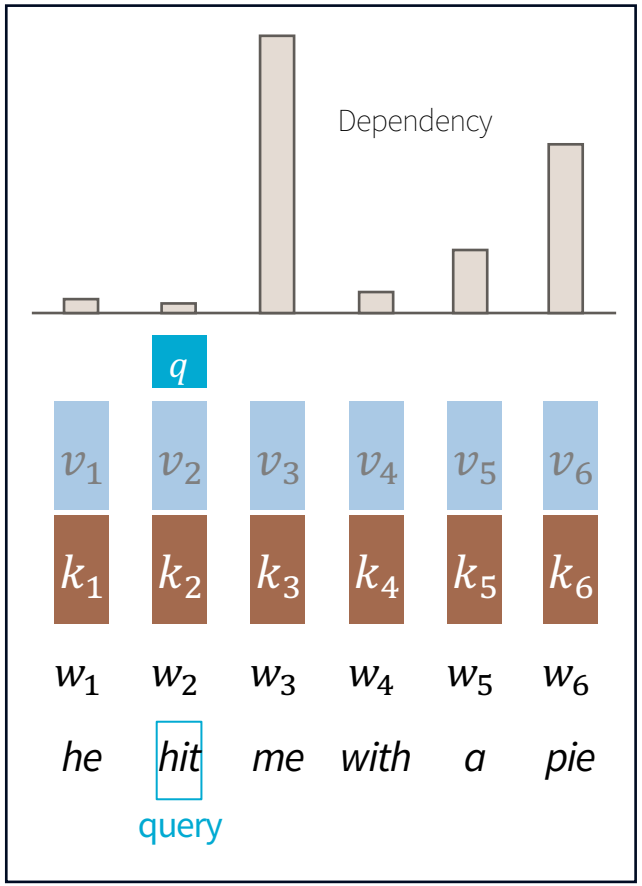
[Transformers]

Let's just talk about multi-head attention.

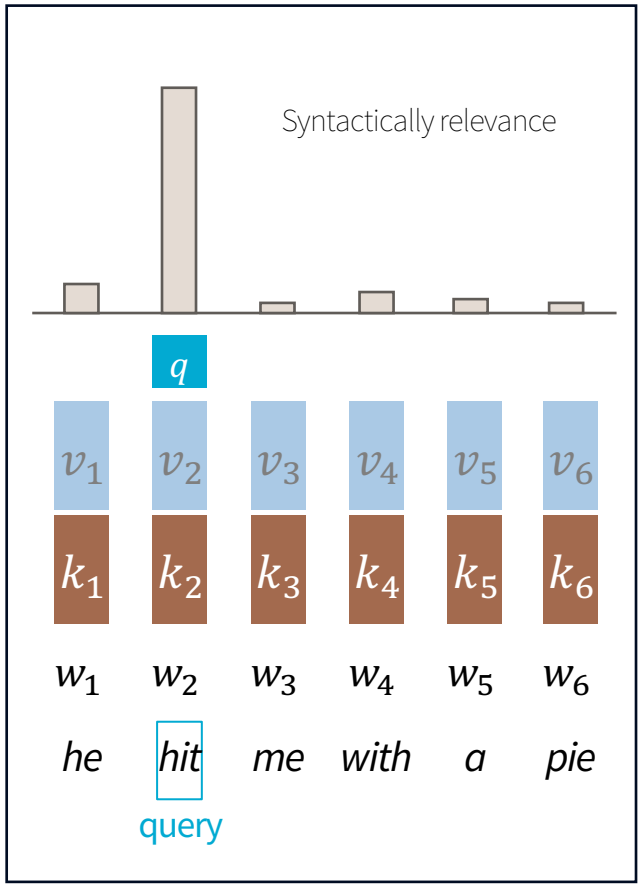
Multi-head Attention

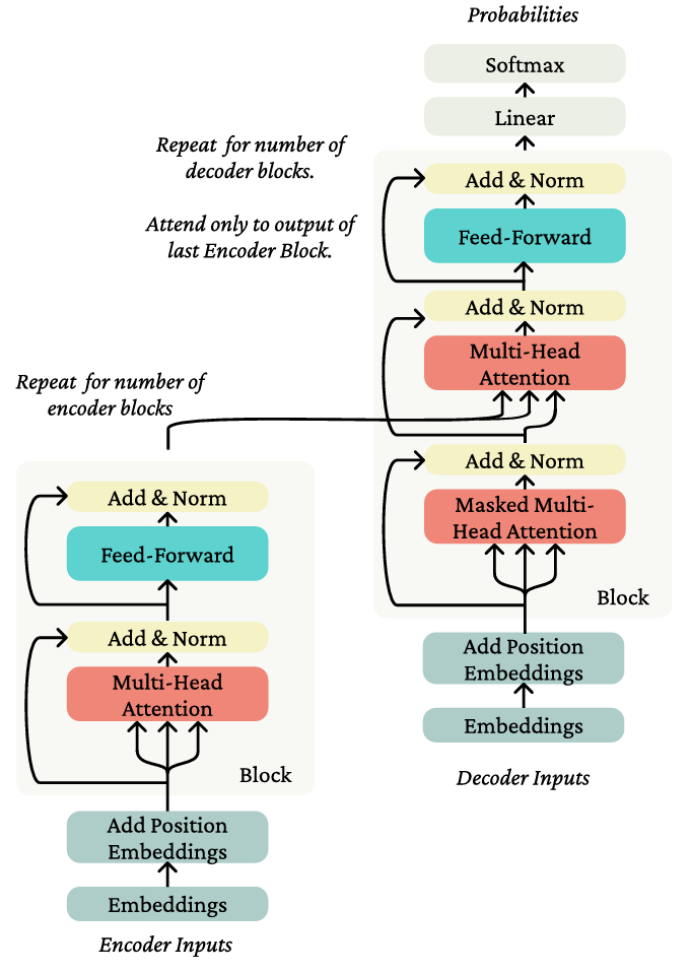
- **Each head gets to "look" at different things, and construct value vectors differently.**

Attention head 1



Attention head 8





[Transformers]

Article Review : 2024-Fall | ASY6001 | **딥러닝**

Transforming Machine Translation:

A Deep Learning System Reaches News Translation Quality Comparable to Human Professionals

P Martin Popel, Marketa Tomkova, Jakub Tomek et al.
(Charles University, University of Oxford, Google Brain)

Nature Communications 11, 4381 (2020)

2024266791 컴퓨터·소프트웨어학과 **윤 예 진**

02
Problem States

Limited Availability of Parallel Data

Expensive Training Large-scale NMT Model

Trade-offs Between Adequacy and Fluency

Problem States for Improving Machine Translation Quality

- **Limited Availability of Parallel Data**
  - Parallel corpora are limited, especially for less commonly paired languages like English and Czech.
  - Poor-quality synthetic data can actually degrade translation performance.

- **Expensive Training Large-scale NMT Model**
  - Large-scale NMT models are expensive to train, especially when using ensembling techniques.

- **Trade-offs Between Adequacy and Fluency**
  - Neural machine translation often improves fluency (making sentences sound natural) but struggles with adequacy (preserving the exact meaning).

  + Traditional evaluations of translation quality are often sentence-by-sentence, without considering context.

Problem States for Improving Machine Translation Quality

- **Limited Availability of Parallel Data**
  - Parallel corpora are limited, especially for less commonly paired languages like English and Czech.
  - Poor-quality synthetic data can actually degrade translation performance.

RQ #1 Can NMT models perform better with limited parallel data by using improved synthetic data generation?

CUBBITT: *Iterated Backtranslation, Block Backtranslation (Block-BT)*

- **Expensive Training Large-scale NMT Model**
  - Large-scale NMT models are expensive to train, especially when using ensembling techniques.

RQ #2 Is there any resource-efficient method to avoid the high computational cost of full ensembling?

CUBBITT: *Checkpoint Averaging*

- **Trade-offs Between Adequacy and Fluency**
  - Neural machine translation often improves fluency (making sentences sound natural) but struggles with adequacy (preserving the exact meaning).
  - + Tradit… considering context.

RQ #3 Does it strike a better balance between fluency and adequacy, prioritizing the preservation of meaning?

CUBBITT: *Context-Aware Evaluation (with Enhanced Adequacy)*

03
Suggestions; CUBBITT

RQ1) Iterated Backtranslation

RQ2) Checkpoint Averaging

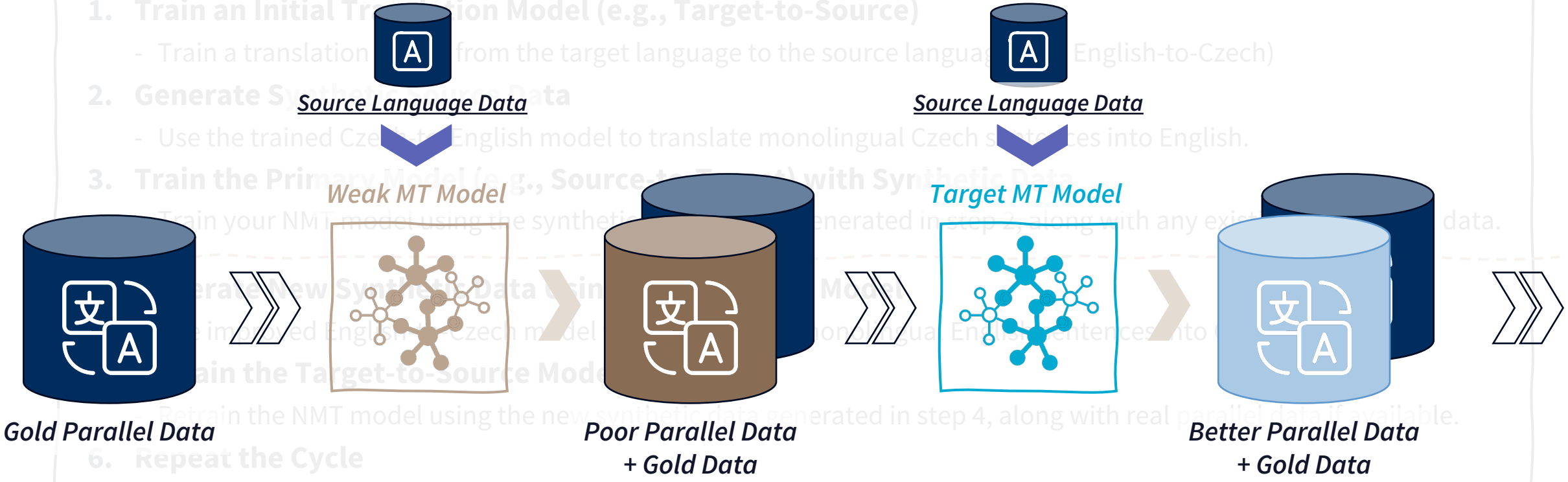
RQ #1 Can We Train NMT Models with Synthetic Parallel Data?

- **Iterated Backtranslation**
  - Backtranslation: create more training data using weak model when we don't have enough parallel data

1. **Train an Initial Translation Model (e.g., Target-to-Source)**
   - Train a translation model from the target language to the source language (e.g. English-to-Czech)
2. **Generate Synthetic Source Data**
   - Use the trained Czech-to-English model to translate monolingual Czech sentences into English.
3. **Train the Primary Model (e.g., Source-to-Target) with Synthetic Data**
   - Train your NMT model using the synthetic parallel data generated in step 2, along with any existing real parallel data.
4. **Generate New Synthetic Data Using the Improved Model**
   - Use improved English-to-Czech model to backtranslate monolingual English sentences into Czech.
5. **Retrain the Target-to-Source Model**
   - Retrain the NMT model using the new synthetic data generated in step 4, along with real parallel data if available.
6. **Repeat the Cycle**
   - Repeat steps 3-5 multiple times, each time using the improved models to generate higher-quality synthetic data.

RQ #1 Can We Train NMT Models with Synthetic Parallel Data?

- **Iterated Backtranslation**
  - Backtranslation: create more training data using weak model when we don't have enough parallel data



RQ #1 Can We Train NMT Models with Synthetic Parallel Data?

- **Iterated Backtranslation**
  - Backtranslation: create more training data using weak model when we don't have enough parallel data

1. **Train an Initial Translation Model (e.g., Target-to-Source)**
   - Train a translation model from the target language to the source language (e.g. English-to-Czech)
2. **Generate Synthetic Source Data** → *UEdin2016*
   - Use the trained Czech-to-English model to translate monolingual Czech sentences into English.
3. **Train the Primary Model (e.g., Source-to-Target) with Synthetic Data**
   - Train your NMT model using the synthetic parallel data generated in **step 2**, along with any existing real parallel data.

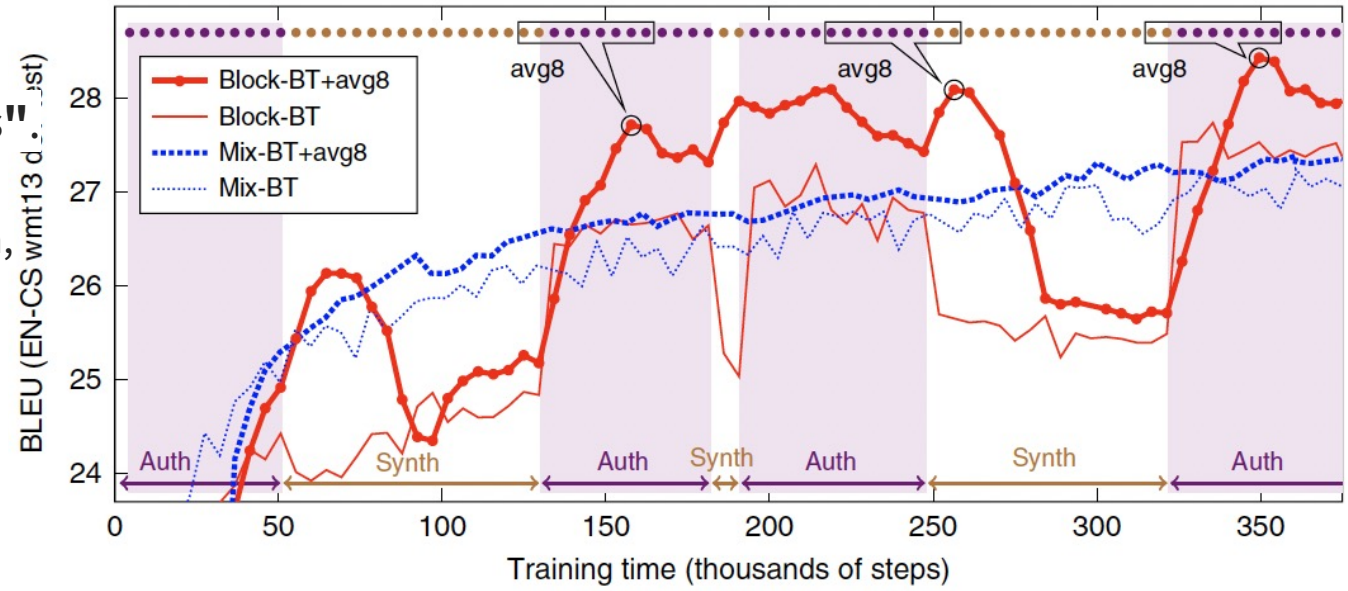
Transformers

4. **Generate New Synthetic Data Using the Improved Model**
   - Use improved English-to-Czech model to backtranslate monolingual English sentences into Czech.
5. **Retrain the Target-to-Source Model**
   - Retrain the NMT model using the new synthetic data generated in **step 4**, along with real parallel data if available.
6. **Repeat the Cycle**
   - Repeat steps 3-5 multiple times, each time using the improved models to generate higher-quality synthetic data.

RQ #1 Can We Train NMT Models with Synthetic Parallel Data?

- **Iterated Backtranslation**
  - Backtranslation: create more training data using weak model when we don't have enough parallel data

- **Block Backtranslation**
  - In standard backtranslation, authentic and synthetic (backtranslated) data are usually mixed together randomly in training.
    - During each training epoch, the model encounters a blend of both data types.
  - **Block backtranslation** divides the training process into **distinct "blocks"**.
    - In each block, the model is trained solely on either authentic or synthetic data, instead of a random mix.



RQ #2 Is There Any Resource-efficient Method to Train MT?

- **Checkpoint Averaging**
  - Checkpoint: created to save the model's progress at intervals during training
  - Instead of using costly ensembling (training multiple models and combining their outputs), CUBBITT uses checkpoint averaging
    - averages the parameters from the last few checkpoints during training. (8 checkpoints)
  - Checkpoint averaging provides a similar performance boost to ensembling but is far more efficient.
    - helps reduce BLEU score variance and stabilizes performance without significantly increasing computational costs.

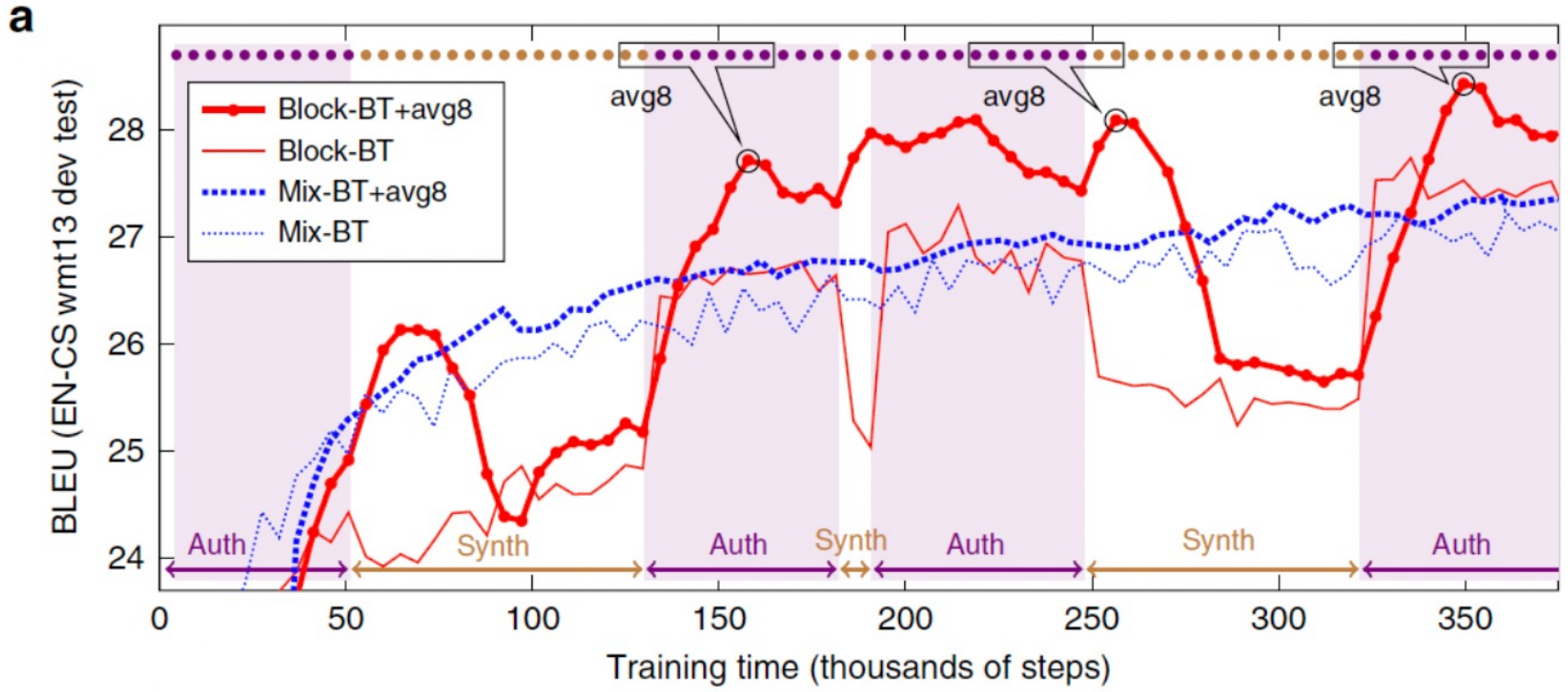
04
Effects

Main Results

Human Evaluations

Main Results (1/2)

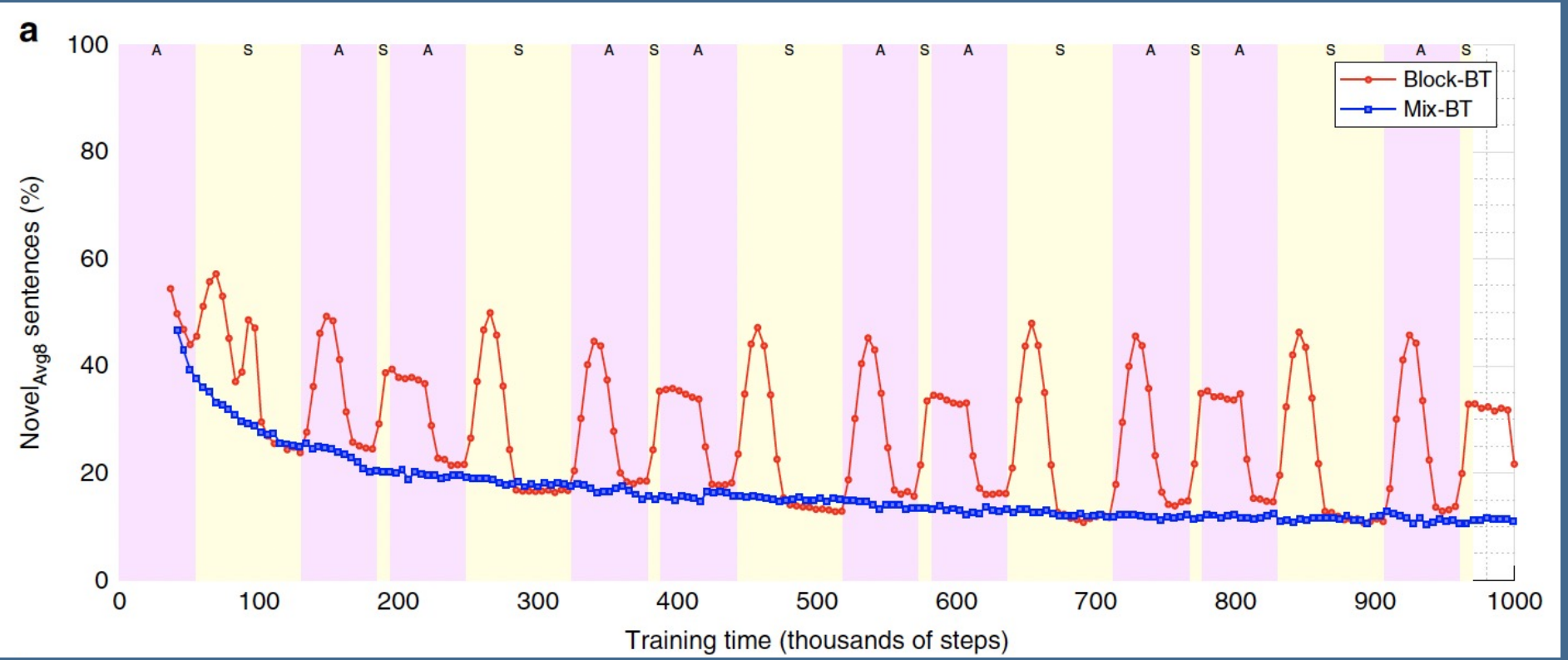
- **How CUBBITT's Backtranslation Technique Contributes to Improving Translation Performance**



- The effect of averaging 8 last checkpoints with block-BT and mix-BT on the translation quality as measured by BLEU on the development set WMT13 newstest.

Main Results (2/3)

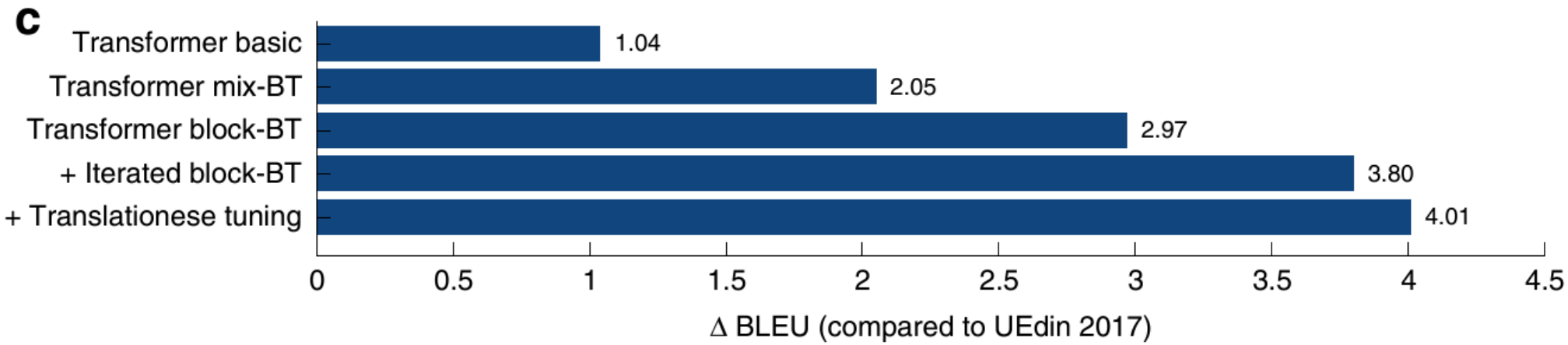
- **Translation Diversity: Block-BT vs. Mix-BT w/ Averaging Checkpoint**



- Block-BT produces a wider range of translations than Mix-BT, suggesting that it is exploring more translation possibilities.
- Novel $_{Avg8}$ sentences: a new translation that is generated during checkpoint averaging

Main Results (3/3)

- **How CUBBITT's Backtranslation Technique Contributes to Improving Translation Performance**

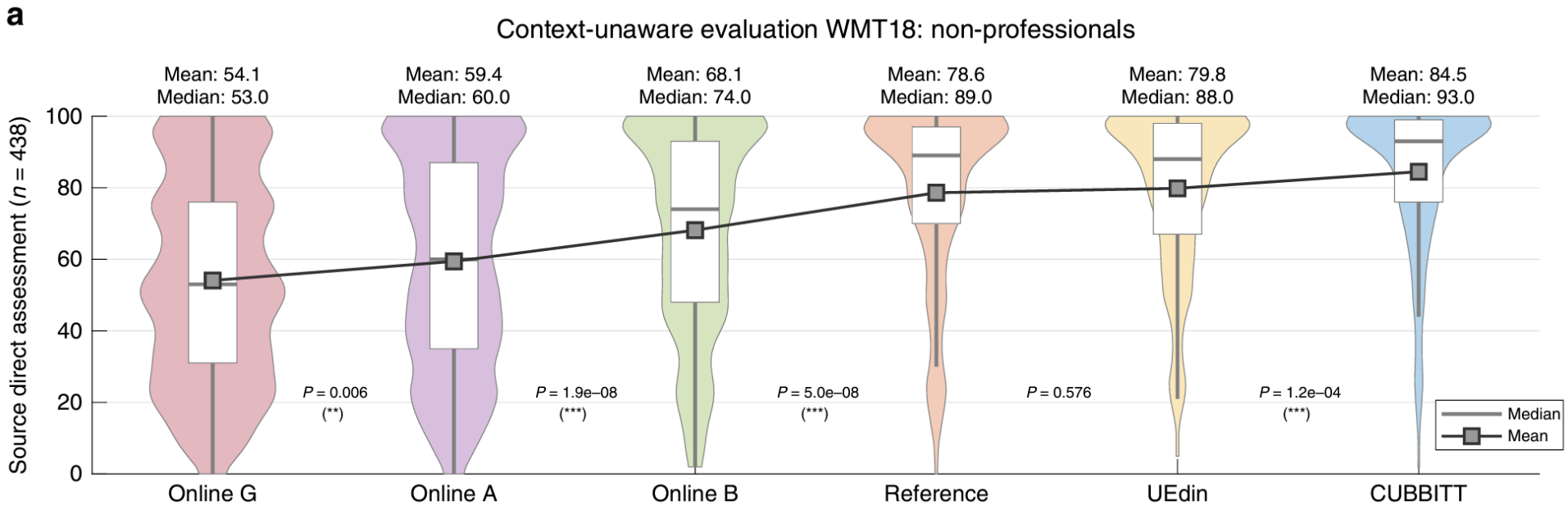


- Starting from the basic Transformer, each additional enhancement (such as Block-BT, iterated backtranslation, and translationese tuning) leads to better performance.

- The final model, with all techniques combined, reaches a BLEU improvement of 4.01 compared to the 2017 state-of-the-art system (UEdin2017), achieving the highest score.

Human Evaluations (1/3)

- **Context-unaware evaluation WMT18 w/ non-professionals**

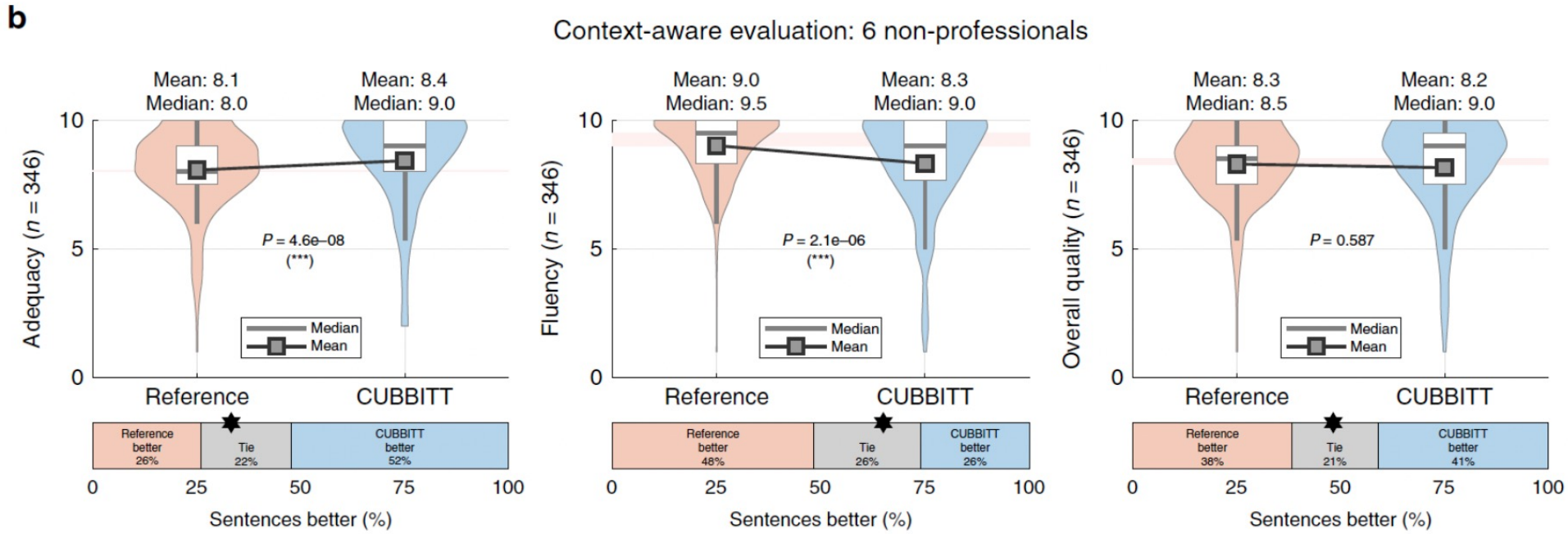


- Showing distributions of source-based direct assessment (SrcDA) of 5 MT systems and human reference translation, sorted by average score.
- CUNI-Transformer outperform other baselines.
- CUNI-Transformer: CUBBITT, Online G/A/B: anonymized online MT systems

Human Evaluations (2/3)

- **Context-aware evaluation WMT18 w/ non-professionals**

  RQ #3 *Does it strike a better balance between fluency and adequacy, prioritizing the preservation of meaning?*

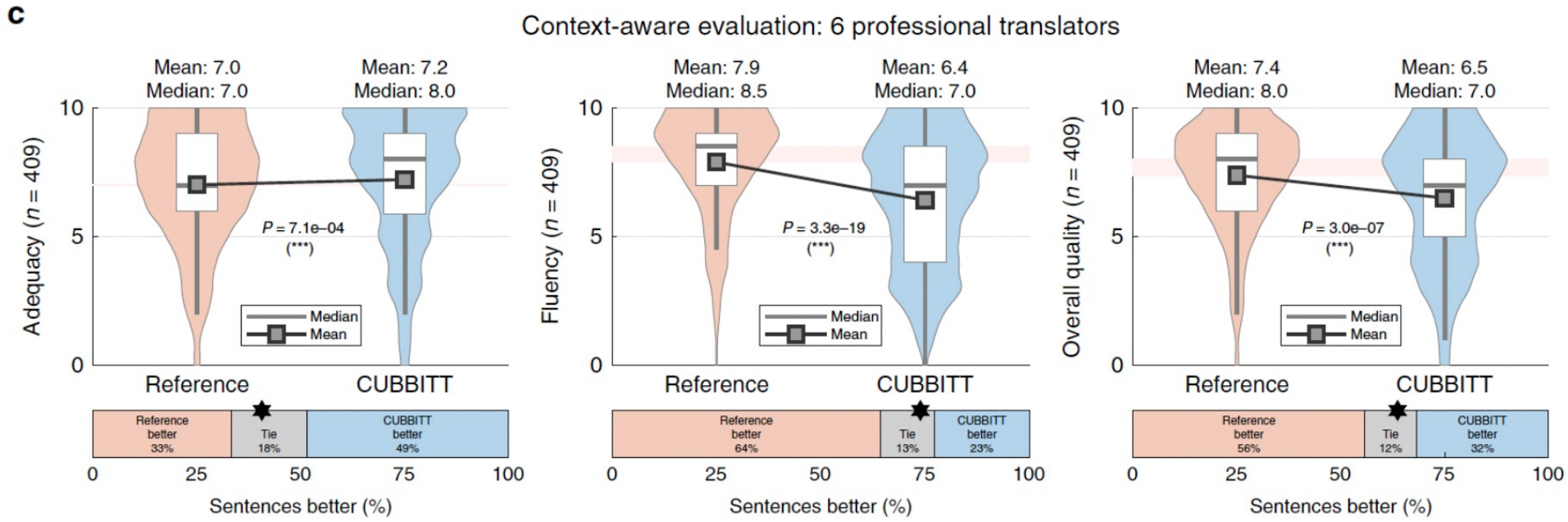


- Adequacy: **CUBBITT** > Human ($p = 4.6\times10^{-8}$)
- Fluency: CUBBITT < **Human** ($p = 2.1\times10^{-6}$)
- Overall Quality: CUBBITT ≥ Human ($p = 0.587$)

Human Evaluations (3/3)

- **Context-aware evaluation WMT18 w/ professionals**

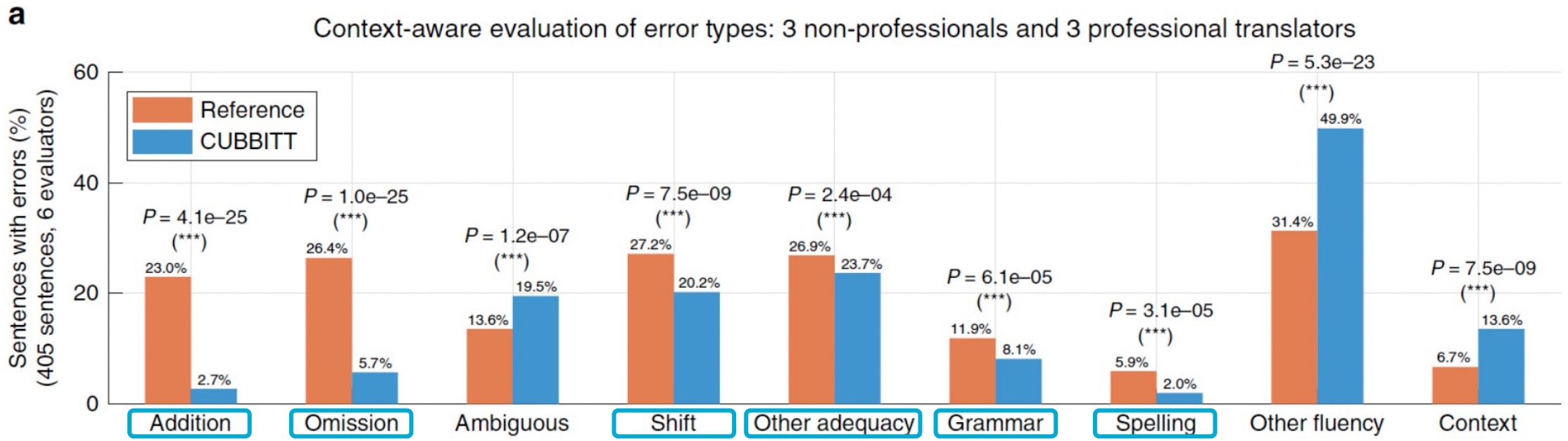
  RQ #3 *Does it strike a better balance between fluency and adequacy, prioritizing the preservation of meaning?*



- Adequacy: **CUBBITT** > Human ($p = 7.1\times10^{-4}$)
- Fluency: CUBBITT < **Human** ($p = 3.3\times10^{-19}$)
- Overall Quality: CUBBITT* < Human ($p = 3.0\times10^{-7}$)

Error Analysis

- **Context-aware evaluation WMT18**



a: Meaning omission by human reference

This event will be his first Scottish appearance since the Aberdeen Asset Management Scottish Open in 2013.

The Czech human translation is fluent, but entirely drops the highlighted word “Scottish”, making the sentence sound like the respective golf player has never played golf publicly since 2013.

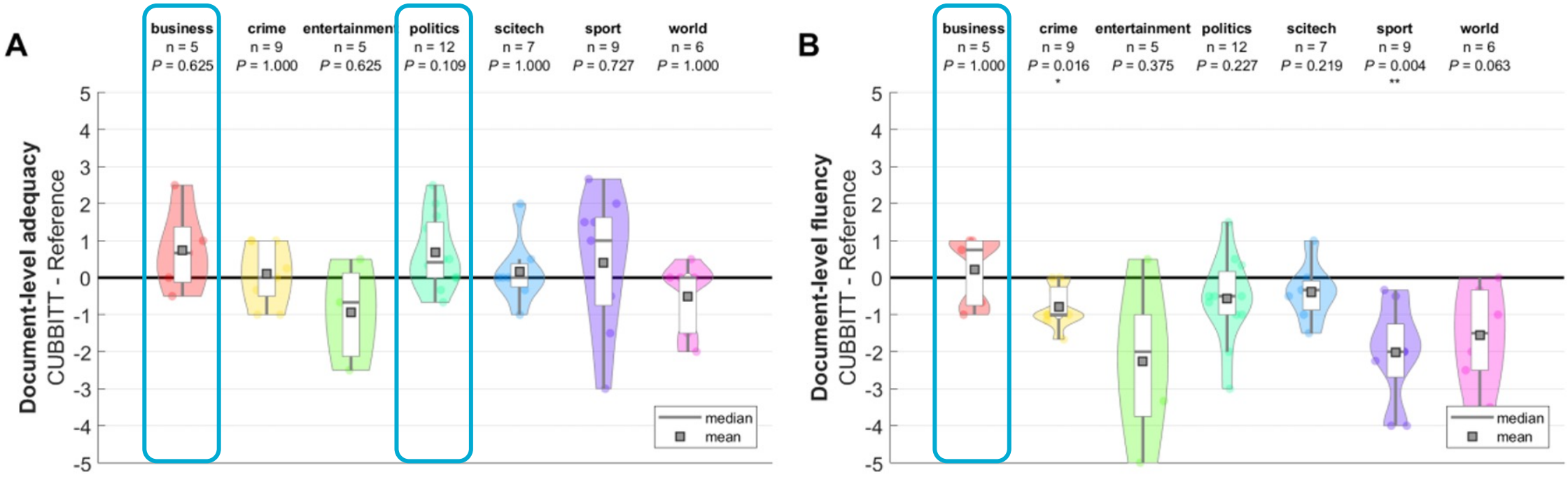
d: Typical CUBBITT error when resolving ambiguous word

And he's got to move up to lightweight [135 pounds] and then we have a fight.

Here, “then we have a fight” (a boxing match) was translated by CUBBITT as “pak se pohádáme“ (then we will quarrel), yielding a funny, but not very adequate translation.

Application

- **Context-aware evaluation of translation quality with respect to document types**



- Business and political news: **CUBBITT** > Human
- Entertainment/Arts & Sports news: CUBBITT < **Human**

05
Conclusion

Summary

- **Machine translation may surpass human translation adequacy under certain conditions, challenging the belief that human translation is the upper bound for quality.**

- **Main Contribution**
  - Introduction of the CUBBITT System:
    - A neural machine translation system that achieves human-level quality in news translation from English to Czech.
    - In terms of meaning preservation (adequacy), CUBBITT's quality surpasses that of professional human translators.

  1. **Block Backtranslation Technique**:
     - CUBBITT utilizes a novel *block backtranslation* (block-BT) regime, which improves training by alternating between authentic and synthetic data blocks. → increases the diversity of training data, leading to better model generalization
  2. **Checkpoint Averaging Synergy**:
     - CUBBITT demonstrates combining block backtranslation with *checkpoint averaging*, which enhances model stability and performance. → smoother & more accurate translations by averaging the last few training checkpoints.
  3. **Context-Aware Evaluation Methodology**:
     - This research accounts for document-level context rather than isolated sentences.

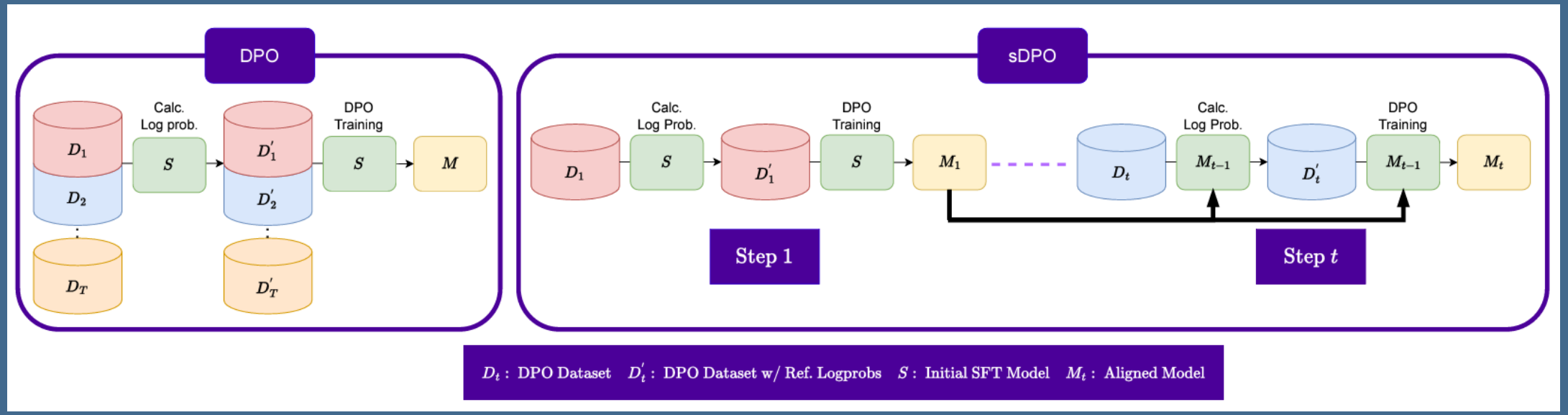
Future Work

- **Limitations**
  1. **Limited Language Pairs and Domains**:
     - While the study primarily focuses on English-to-Czech news translation, the authors acknowledge that CUBBITT's effectiveness across other languages and domains (e.g., technical or literary texts) remains untested.
     - Additional experiments with English-French and English-Polish translations provide some preliminary support but are not exhaustive.
  2. **Dependence on Training Data Quality**:
     - CUBBITT, like other neural machine translation systems, may restrict its applicability in languages or domains with limited training resources.
  3. **Fluency vs. Adequacy Trade-off**:
     - While CUBBITT excels in adequacy, human translations are still rated higher for fluency.
     - Further work is needed to enhance fluency without compromising meaning preservation.
  4. **Contextual Errors in Cross-Sentence Translations**:
     - CUBBITT's translations may include errors when cross-sentence context is necessary.
     - Sometimes it struggles with context-dependent elements like gender and reference consistency across sentences.

Recent Research Trends

Kim et al. (Upstage) “sDPO: Don't Use Your Data All at Once” (arXiv 2024)

- **Stepwise DPO**
  - Preference datasets are divided to be used in multiple steps instead of using them all at once.



Preventing Overoptimization

References

Popel et al. (Charles University) “Transforming Machine Translation: A Deep Learning System Reaches News Translation Quality Comparable to Human Professionals” (Nature Communications 11 2020)

[01 Background]

Vaswani et al. (Google) “Attention Is All You Need” (NIPS2017)

[05 Conclusion]

Kim et al. (Upstage) “sDPO: Don't Use Your Data All at Once” (arXiv 2024)

[Appendix]

Ilya Sutskever, Oriol Vinyals, Quoc V. Le (Google) “Sequence to Sequence Learning with Neural Networks” (NIPS2014)

Thang Luong, Hieu Pham, Christopher D. Manning (Stanford University) “Effective Approaches to Attention-based Neural Machine Translation” (EMNLP2015)

Leo Gao, John Schulman and Jacob Hilton (OpenAI) “Scaling Laws for Reward Model Overoptimization” (PMLR 2024)

Thank You

Yejin Yoon

HYU NLP Lab.
Hanyang University, South Korea

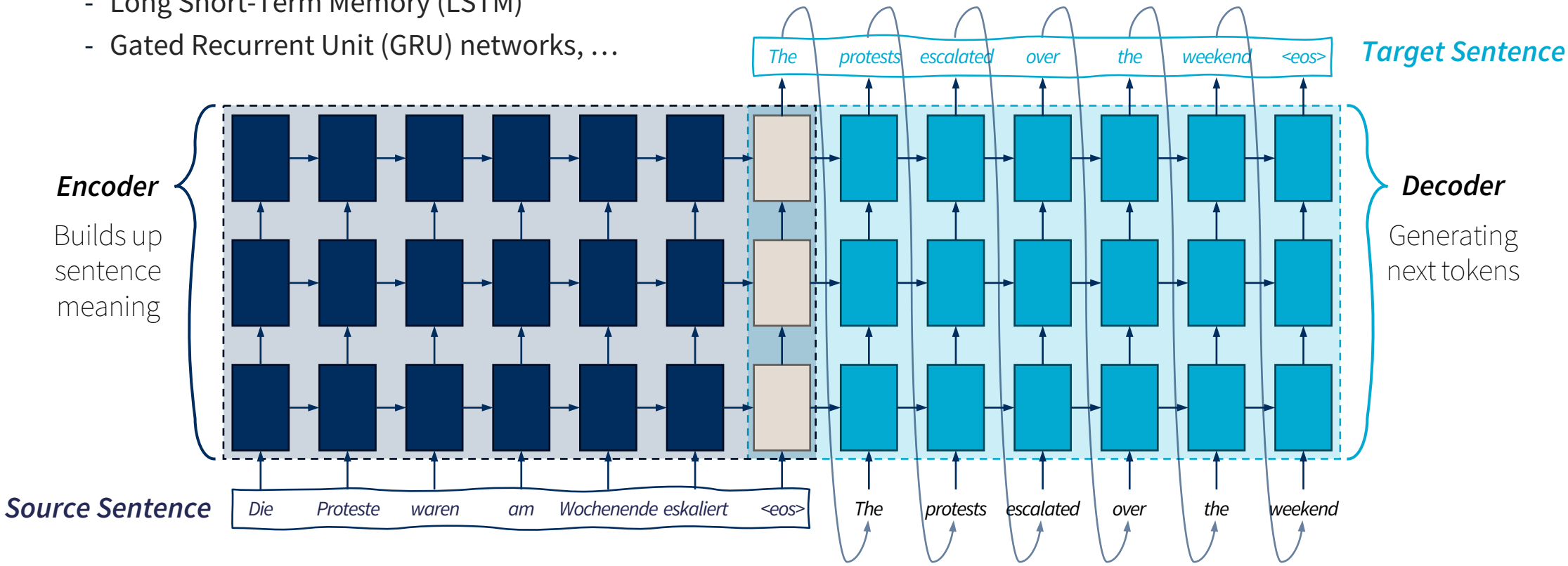
stillwithyou@hanyang.ac.kr

Appendix

A1.1 Brief Overview of Sequence-to-sequence models

Sutskever et al. 2014; Luong et al. 2015

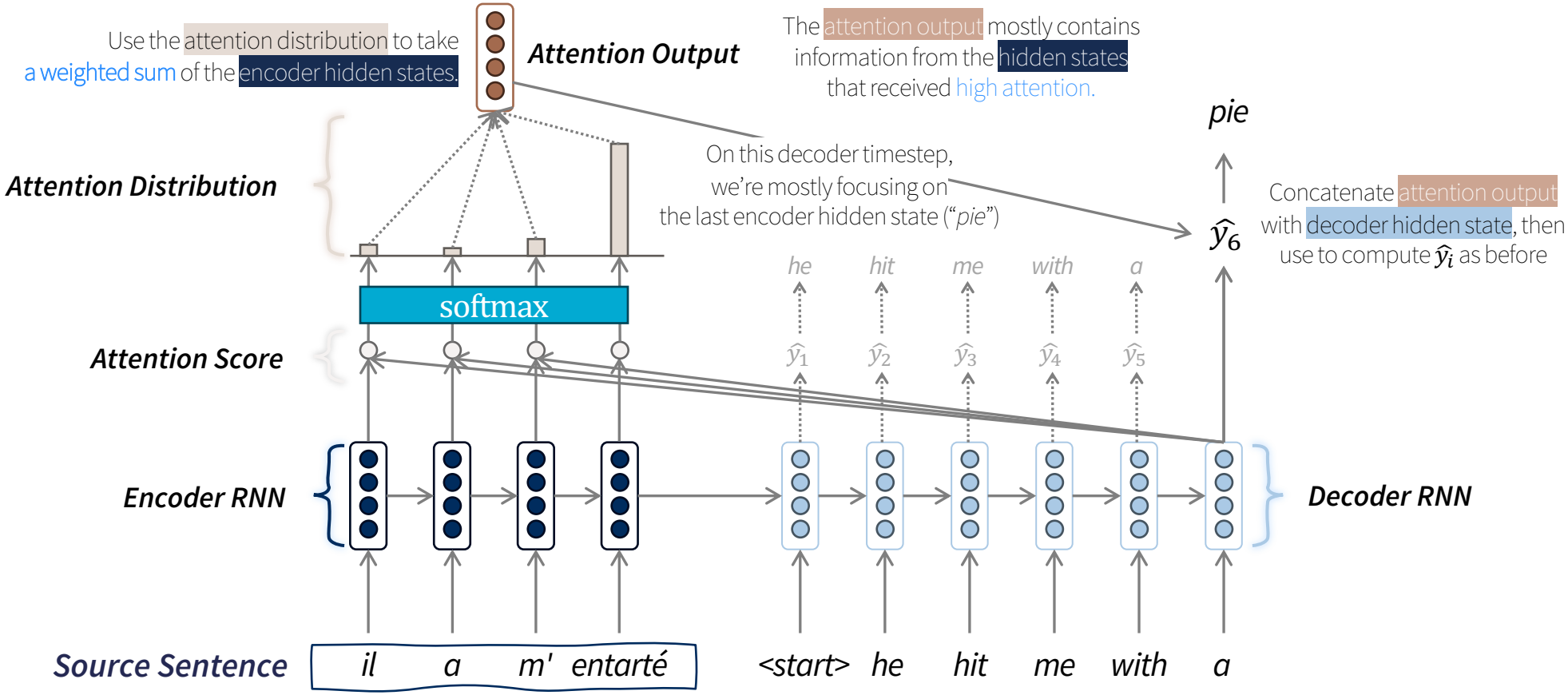
- **Recurrent Neural Networks (RNNs)**
  - RNNs and their variations : SOTA for for Sequence Modeling tasks (e.g., Machine Translation)
    - Long Short-Term Memory (LSTM)
    - Gated Recurrent Unit (GRU) networks, …



The hidden states from RNN layer i are the inputs to RNN layer $i+1$

A1.2 Sequence-to-sequence with Attention (1/2)

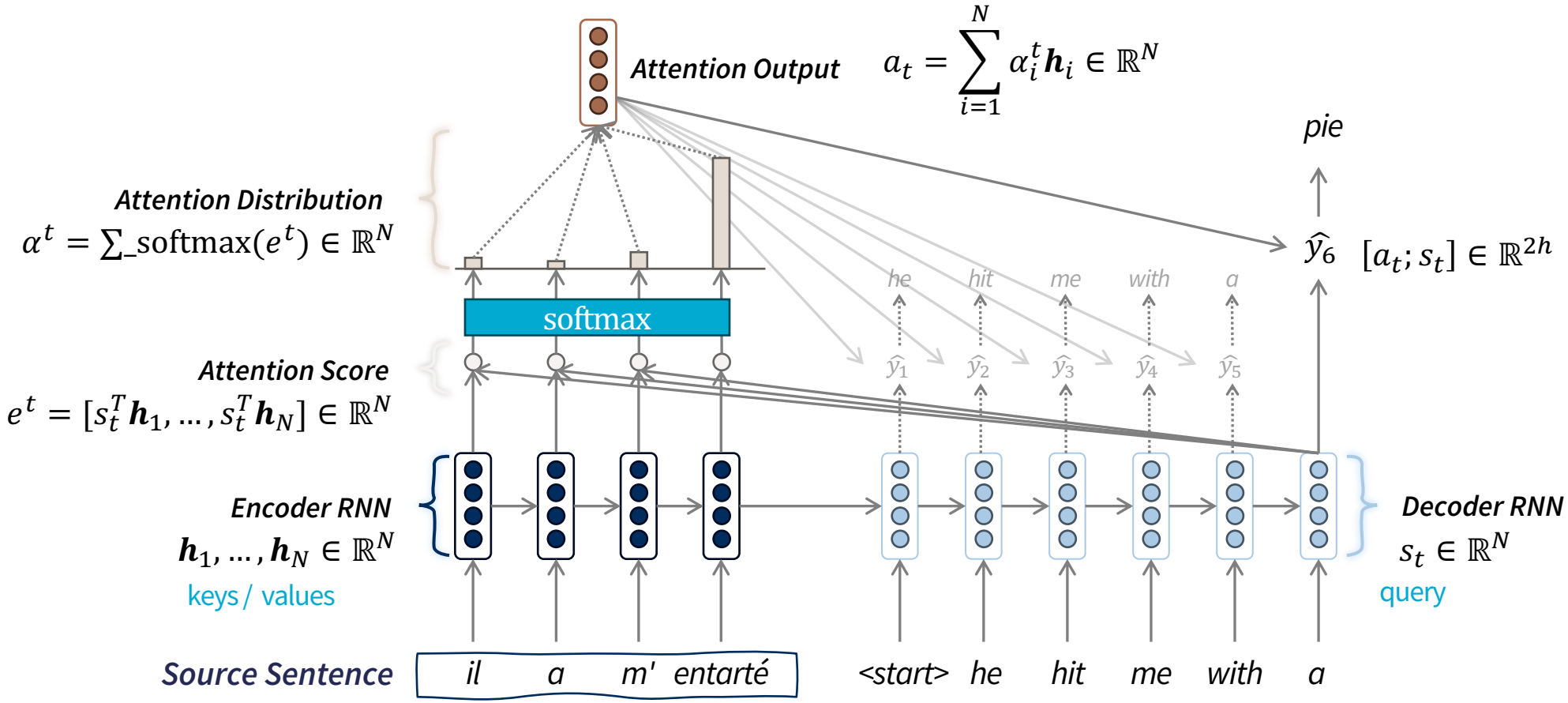
- **Example of Neural Machine Translation**



On each step of the decoder, use *direct connection to the encoder* to focus on a particular part of the source sequence

A1.2 Sequence-to-sequence with Attention (2/2)

- **Example of Neural Machine Translation**



On each step of the decoder, use *direct connection to the encoder* to focus on a particular part of the source sequence

A1.3 Multi-head Attention (1/2)

- **Computationally Efficiency**
  - Even though computing h-head, it's not much costly.
  - Let $X = [x_1; \ldots; x_n] \in \mathbb{R}^{n \times d}$ be the concatenation of input vector.
  - $XQ \in \mathbb{R}^{n \times d}$ and likewise for XK and XV.
  - Output is defined as $\text{output} = \text{softmax}(XQ(XK)^\intercal)XV \in \mathbb{R}^{n \times d}$

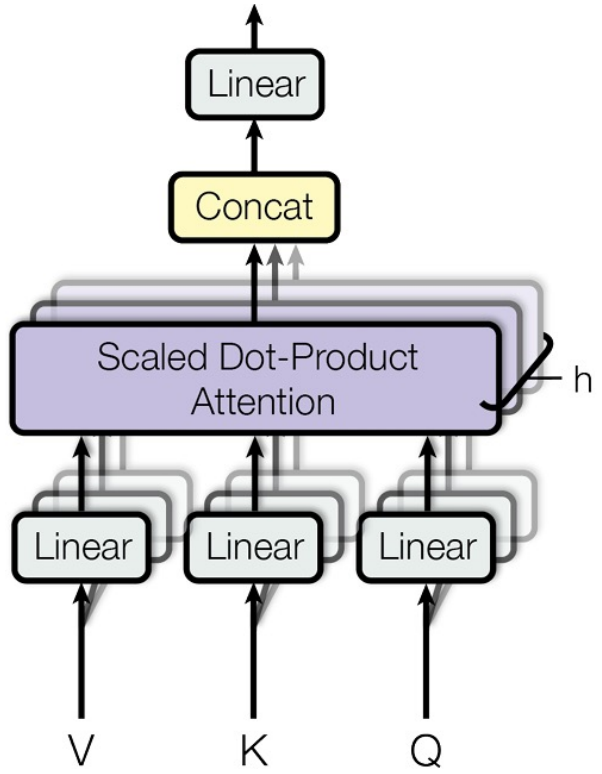
[Attention]

$XQ \quad (XK)^\intercal = XQ(XK)^\intercal \in \mathbb{R}^{n \times n}$

Take the query-key dot products in one matrix multiplication

$\text{softmax}\left(XQ(XK)^\intercal\right) XV = \text{Output} \in \mathbb{R}^{n \times d}$

Take Softmax and compute the weighted average with another matrix multiplication

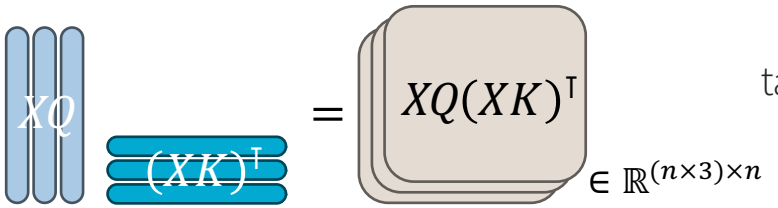


[Multi-Head Attention]

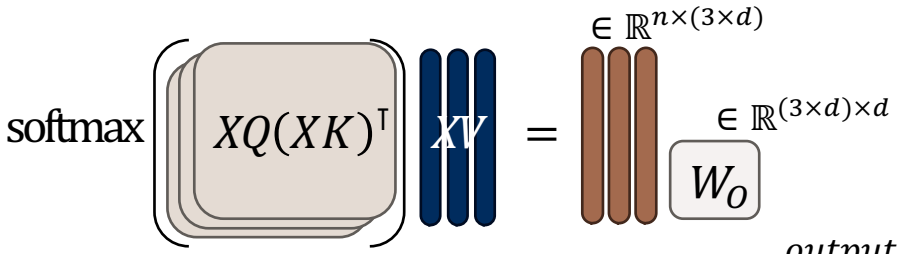
A1.3 Multi-head Attention (2/2)

- **Computationally Efficiency**
  - Even though computing h-head, it's not much costly.
  - Let $X = [x_1; \ldots; x_n] \in \mathbb{R}^{n \times d}$ be the concatenation of input vector.
  - $XQ \in \mathbb{R}^{n \times d}$ and likewise for XK and XV.
  - Output is defined as output $= \text{softmax}(XQ(XK)^\intercal)XV \in \mathbb{R}^{n \times d}$
  - And then reshape XQ , XK and XV to $\mathbb{R}^{n \times h \times d/h}$.

[Multi-head Attention]

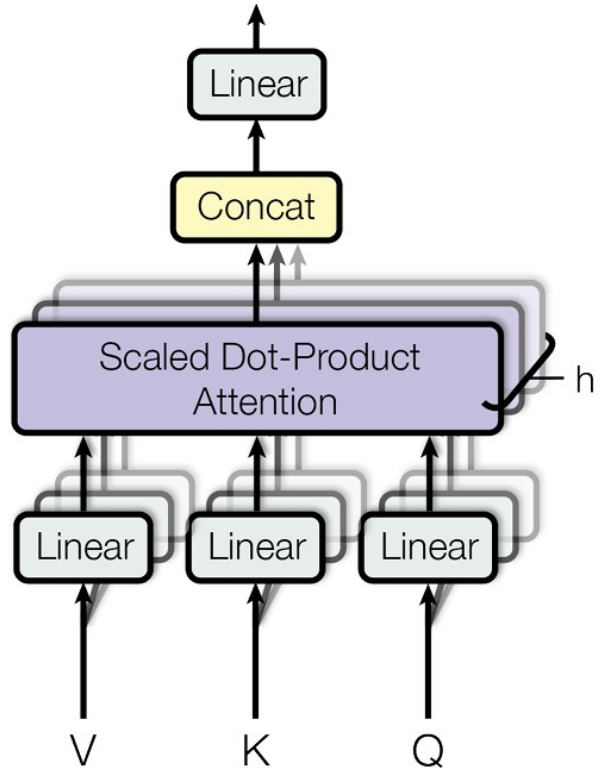


Let $h = 3$,
take the query-key dot products
in one matrix multiplication



Take Softmax and compute
the weighted average
with another matrix multiplication

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$



[Multi-Head Attention]

A5.1 Recent Research Trends

Leo Gao, John Schulman and Jacob Hilton (OpenAI) “Scaling Laws for Reward Model Overoptimization” (PMLR 2024)

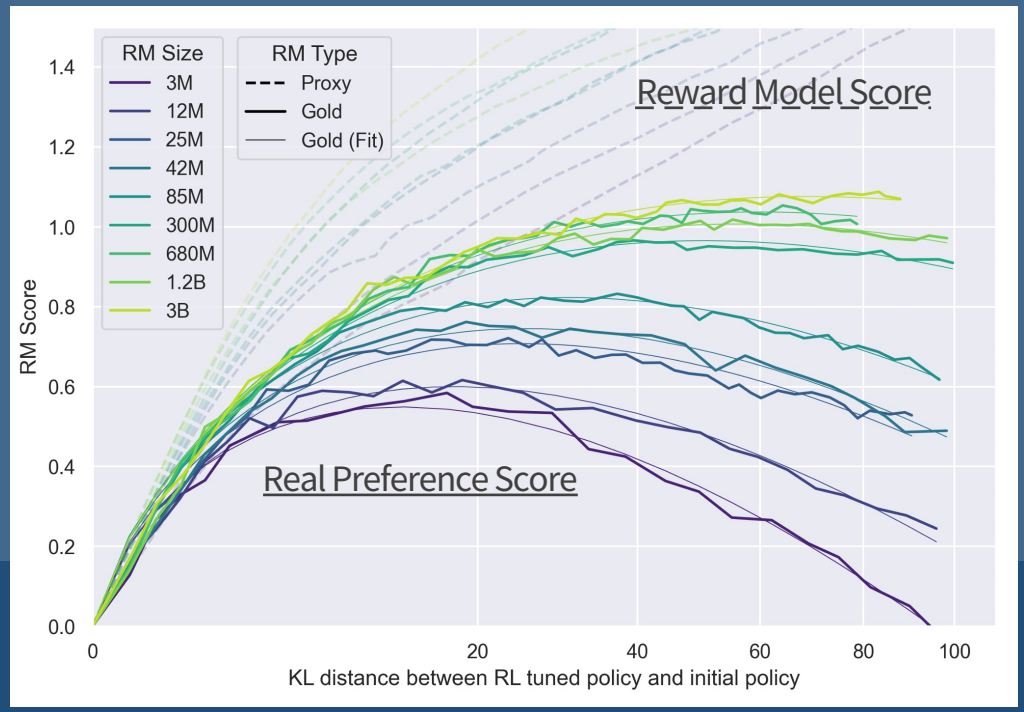
- **Overoptimization**
  - Goodhart’s Law

> “ When a measure becomes a target, it ceases to be a good measure. ”

e.g. In machine learning, this effect arises with **proxy objectives** provided by static learned models, such as discriminators and reward models.

- Solid Line: Gold RM Score = ***Real Preference Score***
- Dash Line: Proxy RM Score

- Optimizing too much against such proxy model, eventually *hinders* the true objective



Reward models don’t fully represent the real world.