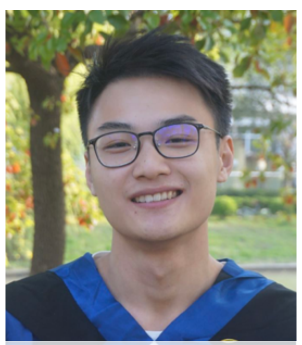
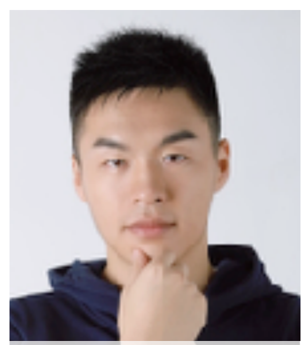
# Neural Machine Translation with Monolingual Translation Memory



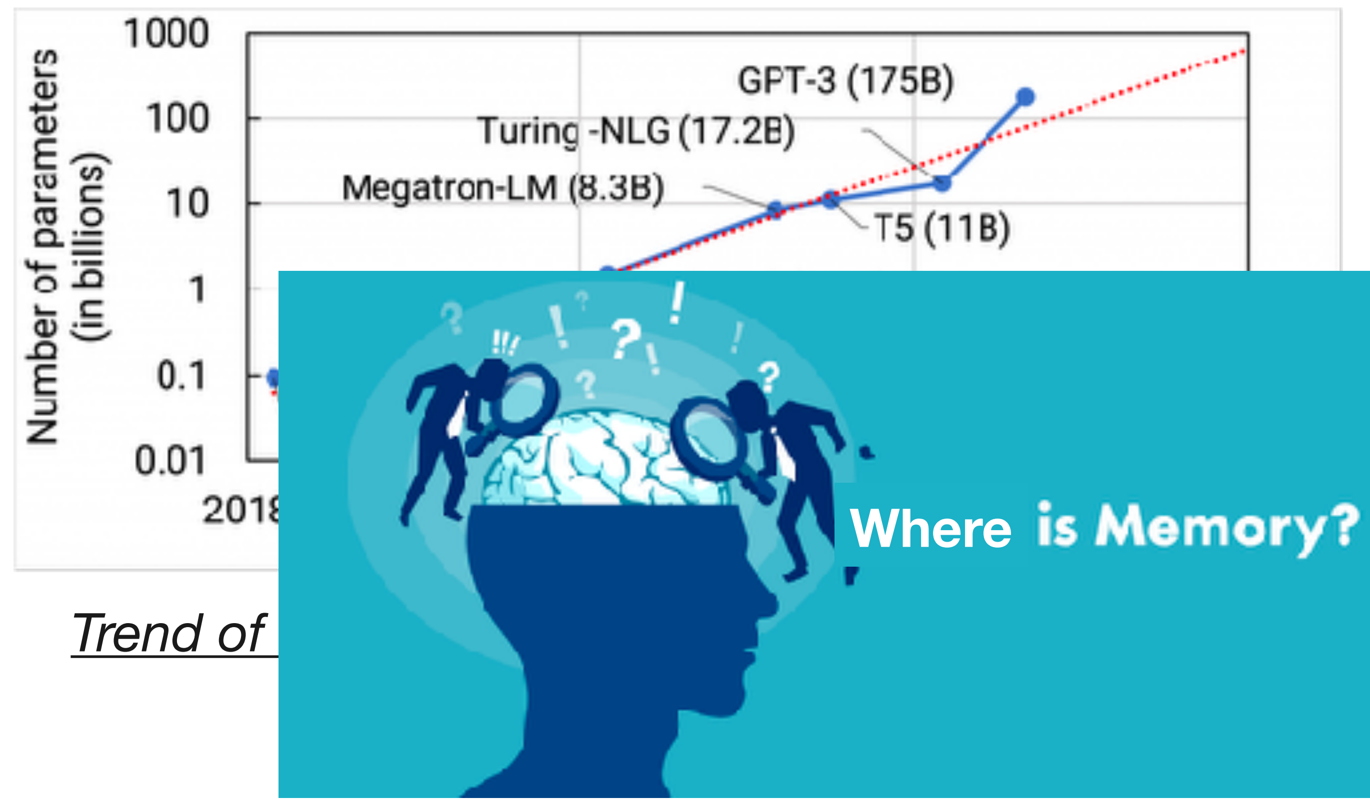
**Deng Cai**[1], Yan Wang[2], Huayang Li[2], Wai Lam[1], Lemao Liu[2]

[1] 香港中文大學 The Chinese University of Hong Kong

[2] Tencent AI Lab

# Background: NLP

knowledge(memory) and
reasoning ability
are mixed in opaque
model parameters

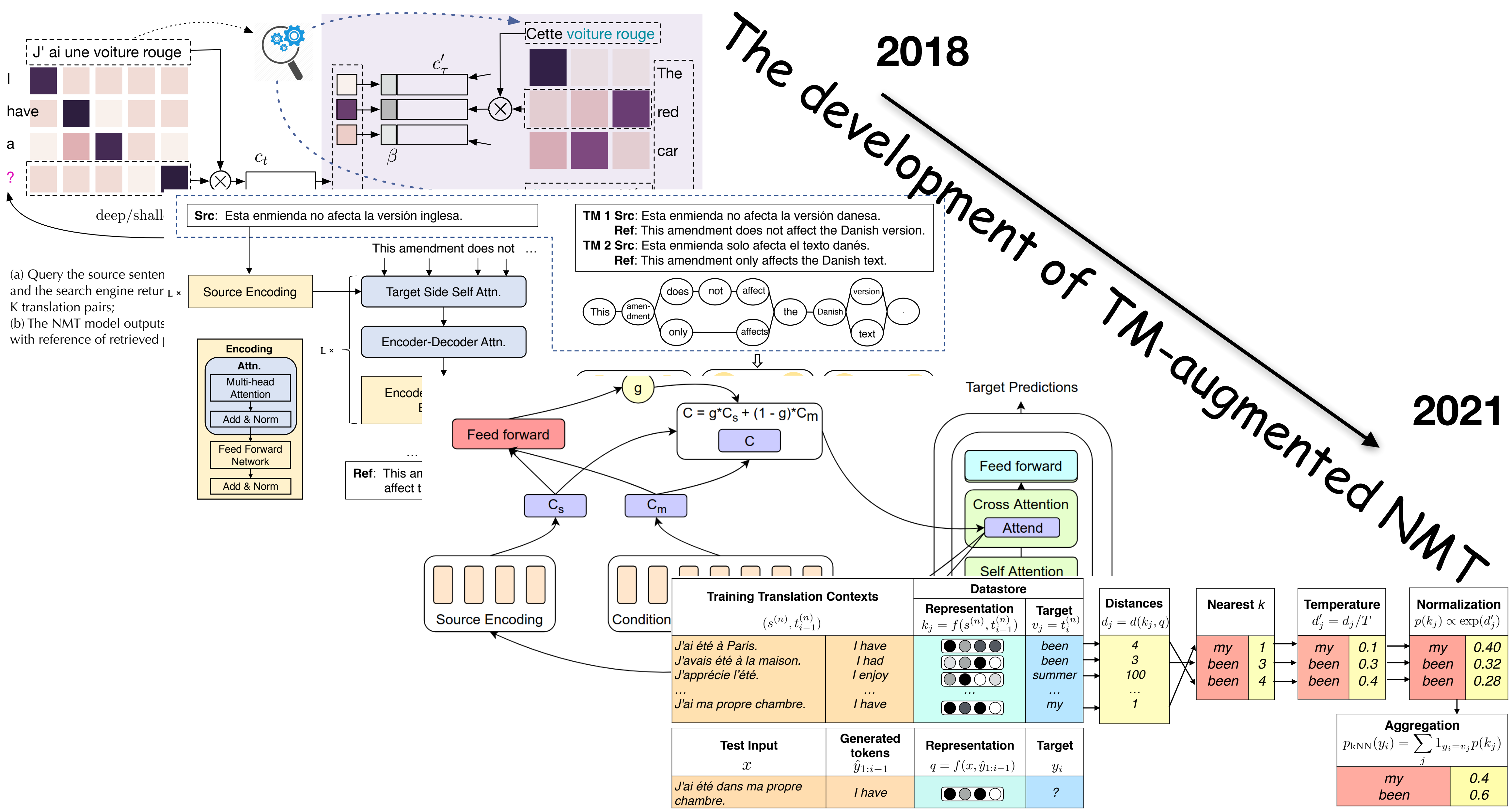



*Trend of*

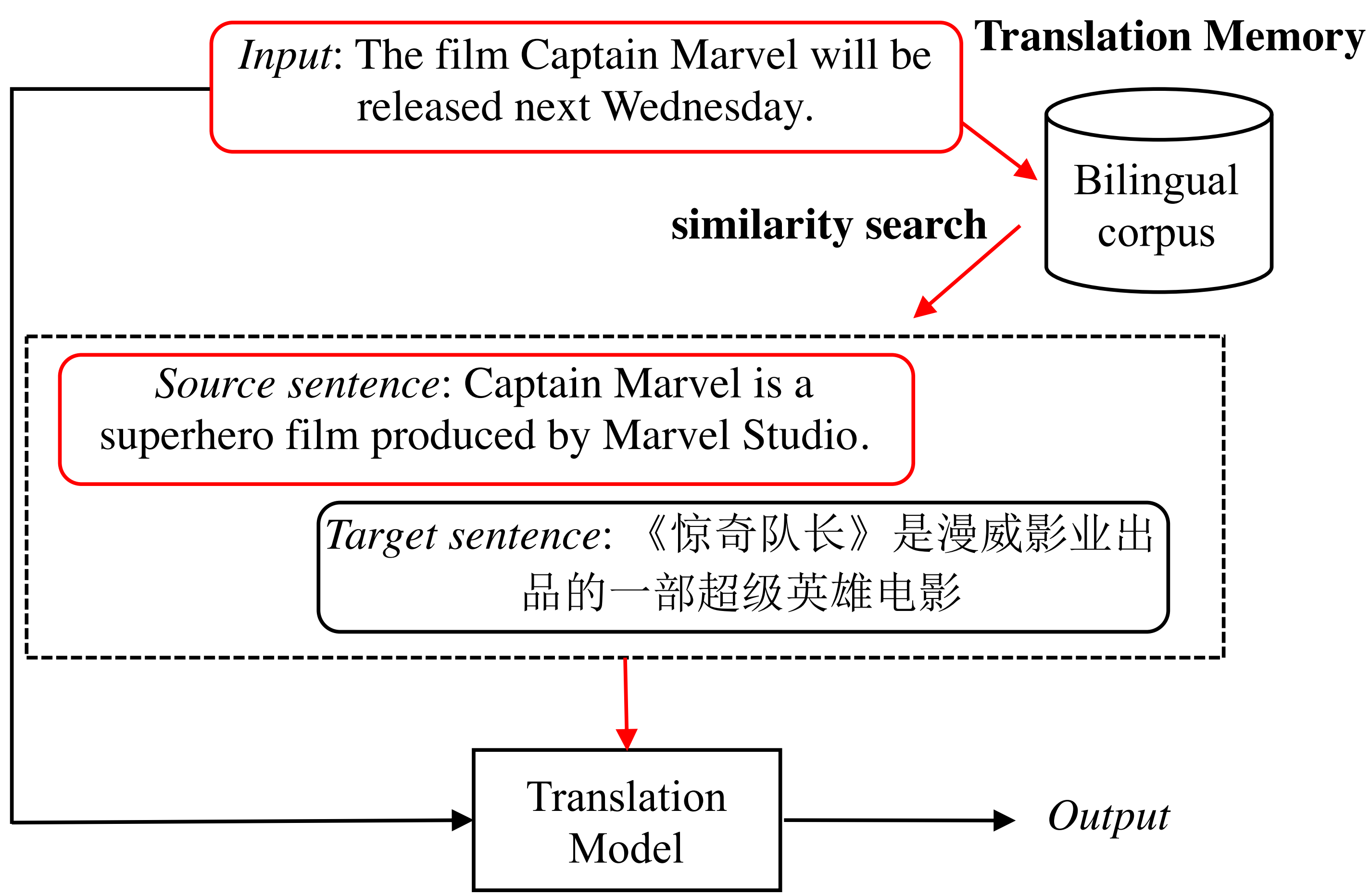
parametric neural networks + non-parametric memory

- transferability: memory can be purposefully changed, expanded or filtered.
- interpretability: influential memory can be manually inspected and interpreted.

# Background: NMT+TM



# Background: NMT+TM




- Traditional TM-augmented NMT framework
    - uses bilingual corpus (training data) as TM
    - employs source (context) similarity search for memory retrieval

# Introduction

- Limitations of the traditional TM-augmented NMT framework
  - **search space**: **bilingual corpus** - source-target translation pairs (training data)
  - **search method**: **heuristic search** - non-learnable, not end-to-end optimized, and lacks for the ability to adapt to specific downstream NMT models
- Our framework
  - **monolingual** memory
  - **learnable & cross-lingual** memory retrieval

# Our Framework

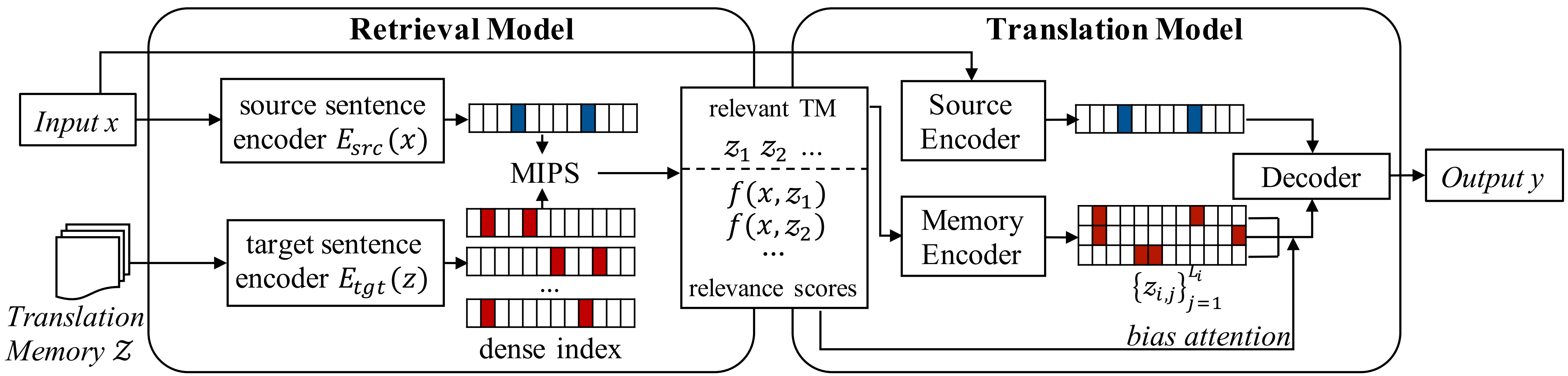
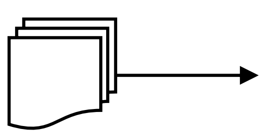



Figure 2: Overall framework. For an input sentence $x$ in the source language, the retrieval model uses Maximum Inner Product Search (MIPS) to find the top-$M$ TM sentences $\{z_i\}_{i=1}^{M}$ in the target language. The translation model takes $\{z_i\}_{i=1}^{M}$ and corresponding relevance scores $\{f(x, z_i)\}_{i=1}^{M}$ as input and generate the translation $y$.

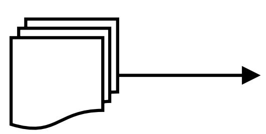
# Our Framework



*Translation Memory* $\mathcal{Z}$

| Source (En) | Translation (De) |
|---|---|
| Criteria for technical regulations | Kriterien für technische Regelungen |
| Contributions to the running costs | Beiträge zu den laufenden Ausgaben |
| … | … |
| Trade with accession countries | Handel mit den Beitrittsländern |

# Our Framework

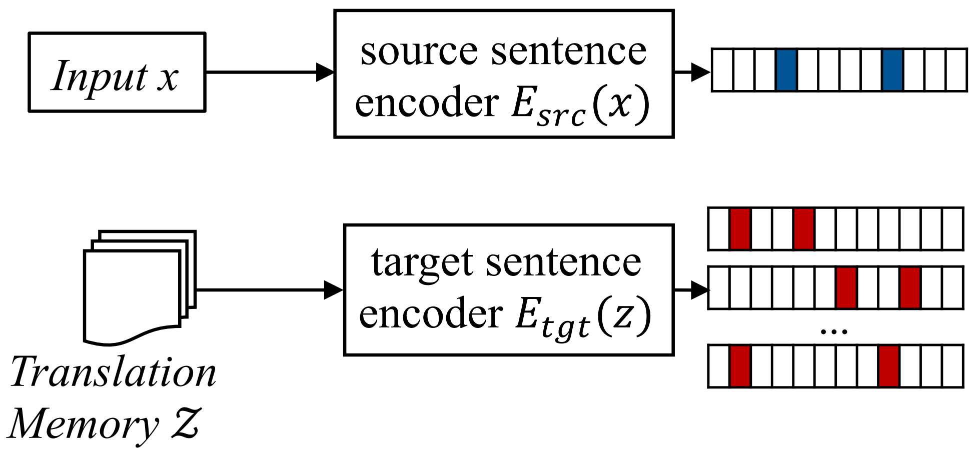


*Translation Memory $\mathcal{Z}$*

| **Translation (De)** |
| --- |
| Kriterien für technische Regelungen |
| Beiträge zu den laufenden Ausgaben |
| ... |
| Handel mit den Beitrittsländern |

\+ **Monolingual data**

# Our Framework



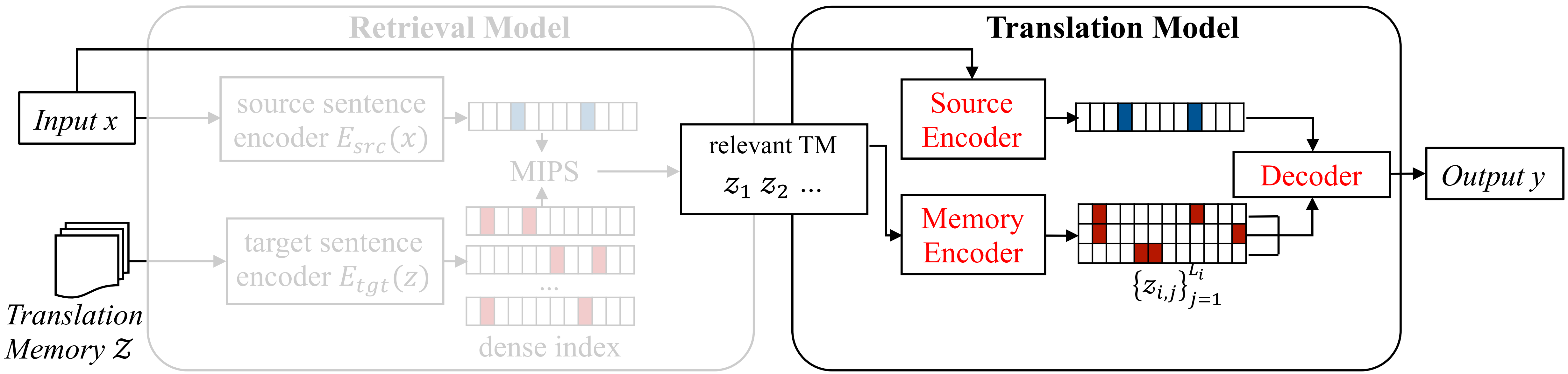

$$E_{\text{src}}(x) = \text{normalize}(W_{\text{src}}\text{Trans}_{\texttt{src}}(x))$$
$$E_{\text{tgt}}(z) = \text{normalize}(W_{\text{tgt}}\text{Trans}_{\texttt{tgt}}(z))$$

$$f(x, z) = E_{\text{src}}(x)^{\text{T}} E_{\text{tgt}}(z)$$

★ **Monolingual Memory**

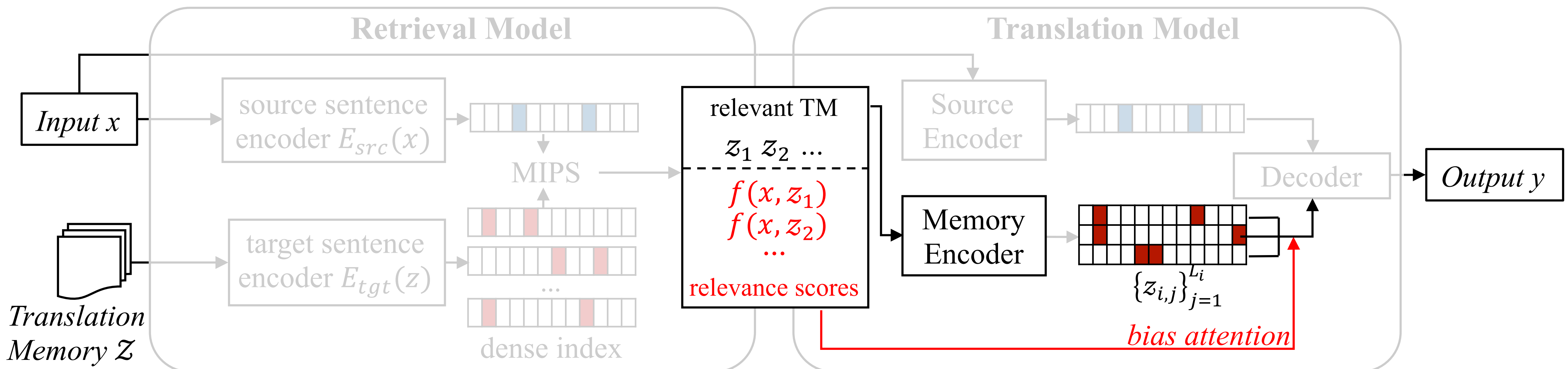
- *directly* connects source-side input and target-side memories.
- *abundant* data in the target language can be used as TM.

# Our Framework




- Changes to standard translation models (Transformer)
  - A separate memory encoder for retrieved TM.
  - The decoder attends over the output of both the source encoder and the memory encoder.

# Our Framework

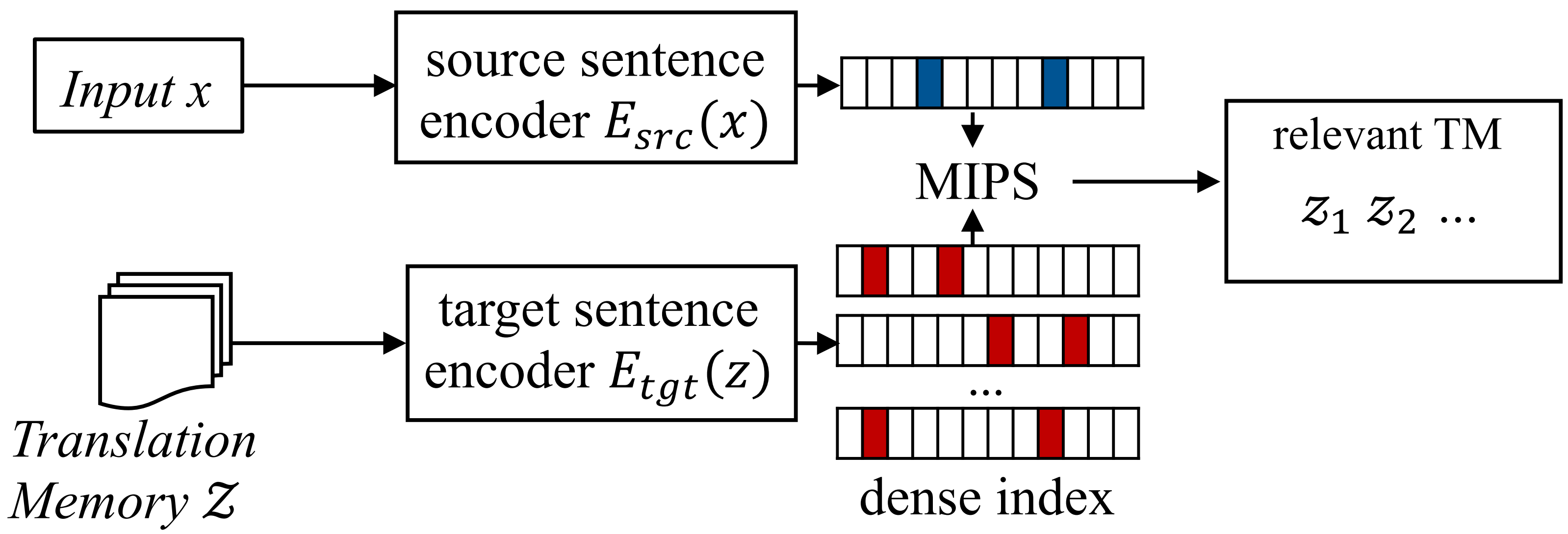



$$\alpha_{ij} = \frac{\exp(h_t{}^{\mathrm{T}} W_m z_{i,j} + \boxed{\beta f(x, z_i)})}{\sum_{i=1}^{M} \sum_{k=1}^{L_i} \exp(h_t{}^{\mathrm{T}} W_m z_{i,k} + \boxed{\beta f(x, z_i)})}$$

★ **Task-Specific Retrieval**

- *unifies* the memory retriever and the downstream NMT model into a *learnable* whole.
- memory retrieval can be *end-to-end* optimized for the *translation* objective.

# Our Framework




★ **Fast Retrieval**

- The selection of the most relevant memories can be reduced to Maximum Inner Product Search (MIPS).
- With off-the-shelf vector search toolkit (FAISS), the search can be made incredibly efficient.

# Experiments

- Conventional Experiments
- Low-resource Scenarios
- Non-parametric Domain Adaptation

# Experiments: Conventional

| Dataset | #Train Pairs | #Dev Pairs | #Test Pairs |
|---|---|---|---|
| **En⇔Es** | 679,088 | 2,533 | 2,596 |
| **En⇔De** | 699,569 | 2,454 | 2,483 |

Table 1: Data statistics for the JRC-Acquis corpus.

| # | System | Retriever | Es⇒En | | En⇒Es | | De⇒En | | En⇒De | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Dev | Test | Dev | Test | Dev | Test | Dev | Test |
| | | | *Existing NMT systems** | | | | | | | |
| | Gu et al. (2018) | source similarity | 63.16 | 62.94 | - | - | - | - | - | - |
| | Zhang et al. (2018) | source similarity | 63.97 | 64.30 | 61.50 | 61.56 | 60.10 | 60.26 | 55.54 | 55.14 |
| | Xia et al. (2019) | source similarity | 66.37 | 66.21 | 62.50 | 62.76 | 61.85 | 61.72 | 57.43 | 56.88 |
| | | | *Our NMT systems* | | | | | | | |
| 1 | *this work* | None | 64.25 | 64.07 | 62.27 | 61.54 | 59.82 | 60.76 | 55.01 | 54.90 |
| 2 | | source similarity | 66.98 | 66.48 | 63.04 | 62.76 | 63.62 | 63.85 | 57.88 | 57.53 |
| 3 | | cross-lingual (fixed) | 66.68 | 66.24 | 63.06 | 62.73 | 63.25 | 63.06 | 57.61 | 56.97 |
| 4 | | cross-lingual (fixed $E_{\text{tgt}}$)† | 67.66 | 67.16 | 63.73 | 63.22 | 64.39 | 64.01 | 58.12 | 57.92 |
| 5 | | cross-lingual† | **67.73** | **67.42** | **64.18** | **63.86** | **64.48** | **64.62** | **58.77** | **58.42** |

Table 2: Experimental results (BLEU scores) on four translation tasks. *Results are from Xia et al. (2019). †The two variants of our method (model #4 and model #5) are significantly better than other baselines with $p$-value $<$ 0.01, tested by bootstrap re-sampling (Koehn, 2004).

- Significant improvements over non-TM NMT model, even outperforming previous bilingual TM-augmented baselines.

# Experiments: Low-resource

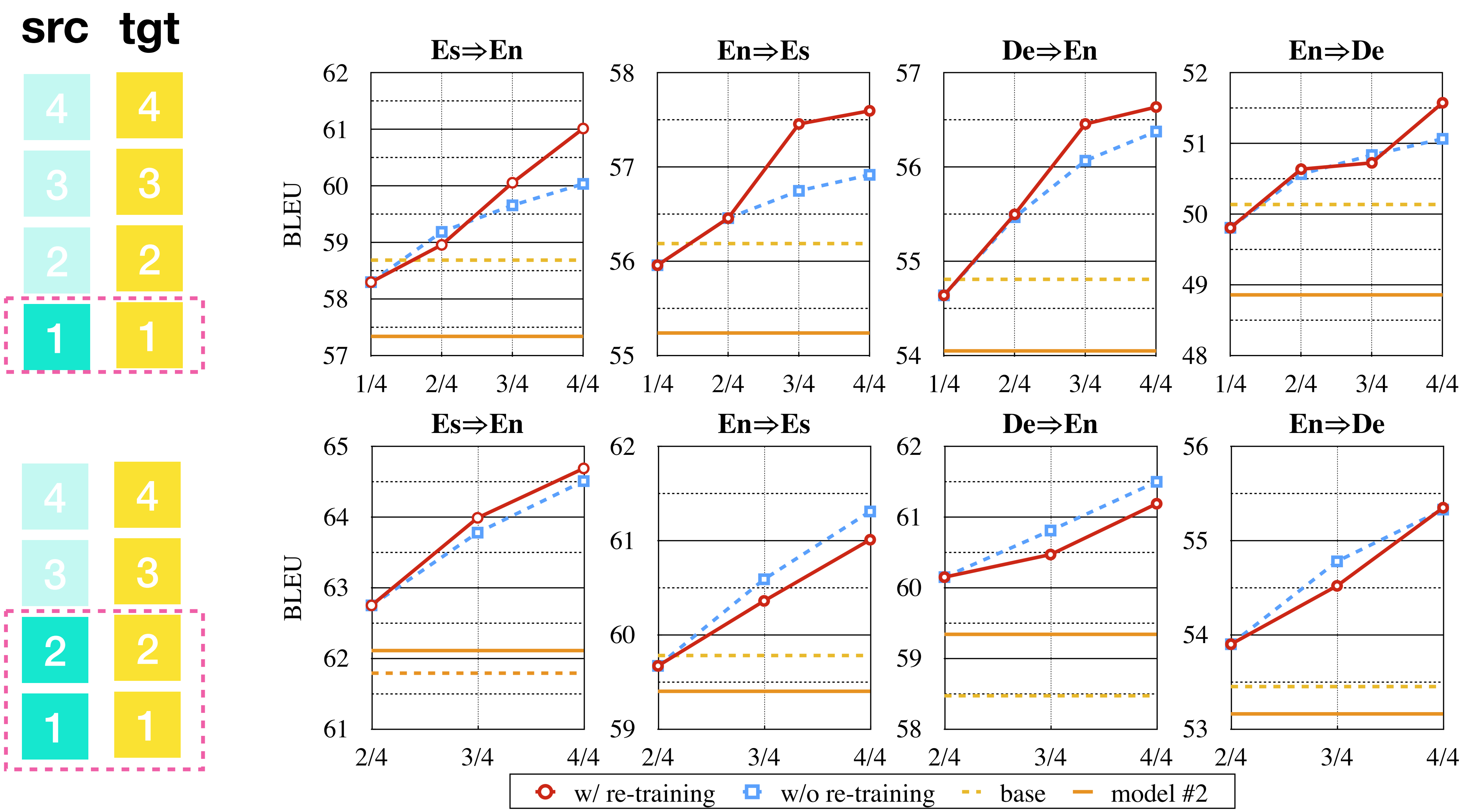



Figure 2: Test results with 1/4 bilingual pairs (upper) and 2/4 bilingual pairs (lower) across different TM sizes.

- Substantial translation quality boost in low-resource scenarios by utilizing more monolingual TM (even without re-training).

# Experiments: Low-resource

## back-translation


| Data | Model | Es⇒En | | En⇒Es | | De⇒En | | En⇒De | |
|---|---|---|---|---|---|---|---|---|---|
| | | dev | test | dev | test | dev | test | dev | test |
| 1/4 bilingual + 4/4 monolingual | Ours | 61.46 | 61.02 | 57.86 | 57.40 | 56.77 | 56.54 | 51.11 | 51.58 |
| | BT | 62.47 | 61.99 | 60.28 | 59.59 | 57.75 | 58.20 | 52.47 | 52.96 |
| | Ours+BT | **65.98** | **65.51** | **62.48** | **62.22** | **62.22** | **61.79** | **56.75** | **56.50** |
| 2/4 bilingual + 4/4 monolingual | Ours | 65.17 | 64.69 | 61.31 | 61.01 | 61.43 | 61.19 | 55.55 | 55.35 |
| | BT | 63.82 | 63.10 | 61.59 | 60.83 | 59.17 | 59.26 | 54.18 | 54.29 |
| | Ours+BT | **66.95** | **66.38** | **63.22** | **62.90** | **63.68** | **63.10** | **57.69** | **57.40** |

- Our method is complementary to back-translation in leveraging additional target-side monolingual corpus.

# Experiments: Domain Adaptation

| | **Medical** | **Law** | **IT** | **Koran** | **Subtitle** | **Avg.** | **Avg. Δ** |
|---|---|---|---|---|---|---|---|
| #Bilingual Pairs | 61,388 | 114,930 | 55,060 | 4,458 | 124,992 | - | - |
| #Monolingual Sents | 184,165 | 344,791 | 165,181 | 13,375 | 374,977 | - | - |
| Using Bilingual Pairs Only | | | | | | | |
| Transformer Base | 47.81 | 51.40 | 33.90 | 14.64 | 21.64 | 33.88 | - |
| Ours | 47.52 | 51.17 | 34.64 | 15.49 | 22.66 | 34.30 | +0.42 |
| + Monolingual Memory | | | | | | | |
| Ours + domain-specific | **50.32** | 53.97 | **35.33** | **16.26** | **22.78** | **35.73** | **+1.85** |
| Ours + all-domains | 50.23 | **54.12** | 35.24 | 16.24 | **22.78** | 35.72 | +1.84 |

Table 4: Test results on domain adaptation.

- Strong cross-domain transferability by hot-swapping domain-specific monolingual TM.

# Summary

★ **Monolingual Memory:** abundant data in the target language can be used as TM

★ **Task-Specific Retrieval:** memory retrieval can be end-to-end optimized for the translation objective

# Summary

- Significant improvements over non-TM NMT model, even outperforming previous bilingual TM-augmented baselines.
- Substantial translation quality boost in low-resource scenarios by utilizing more monolingual TM. (work w/ back-translation)
- Strong cross-domain transferability by hot-swapping domain-specific monolingual TM.

**Github repo:**
[https://github.com/jcyk/copyisallyouneed](https://github.com/jcyk/copyisallyouneed)

# Questions?

Neural Machine Translation with Monolingual Translation Memory

**Deng Cai**, Yan Wang, Huayang Li, Wai Lam, Lemao Liu