FACULTY OF ARTS AND PHILOSOPHY

DEPARTMENT OF TRANSLATION, INTERPRETING AND COMMUNICATION
LT3 LANGUAGE AND TRANSLATION TECHNOLOGY TEAM

Improving Fuzzy Match Augmented Neural Machine Translation through Synthetic Data

Authors: Arda Tezcan, Alina Skidanova, Thomas Moerman

Tezcan, A., Skidanova, A., & Moerman, T. (2024). Improving Fuzzy Match Augmented Neural Machine Translation in Specialised Domains through Synthetic Data. *The Prague Bulletin of Mathematical Linguistics*, *122*, 9–42.

GHENT UNIVERSITY

language and translation technology team

Network of Interdisciplinary Translation Studies
NITS
2025

1. Introduction & Motivation
2. Background: Fuzzy Matches, Back-Translation & NMT
3. Proposed Method: Combining Back-Translation & Neural Fuzzy Repair
4. Experimental Setup
5. Results & Analysis
6. Conclusion & Future Work

GHENT UNIVERSITY

language and translation technology team

Network of Interdisciplinary Translation Studies NITS 2025

Intro & Motivation

Intro & Motivation

- **While powerful, NMT (and LLMs) can struggle with specialized domains (legal, technical, …) due to limited in-domain parallel data** (Jiao et al., 2023)
- **LLMs are increasingly used for MT, thanks for additional capabilities over NMT models:**
  - Instruction following
  - In-context learning (i.e. few-shot learning)
- **However, LLMs do not guarantee superior performance over specialized NMT models in domain-specific scenarios** (Wassie et al., 2025; Higashiyama, 2024)

Intro & Motivation

- **Fuzzy Match (FM) Augmentation:**
  - Leveraging translations of similar sentences (FMs) from translation memories or training data improves quality (Bulté & Tezcan, 2019; Xu et al., 2020)
  - Challenge: FM effectiveness often depends on *large* bilingual datasets to find *high-similarity* FMs
- **This Study's Goal:**
  - Can we improve FM-augmented NMT when in-domain parallel data is *limited*?
  - Hypothesis: Yes, by leveraging *additional monolingual data* to generate *synthetic parallel data* specifically for FM augmentation in specialized domains

Background

Background: Back-Translation

Data Sources in Specialized Domains

Limited Parallel Data

300K pairs

High-quality human translations
In-domain (Legal/Technical)

HIGH QUALITY

Abundant Monolingual

1.5M sentences

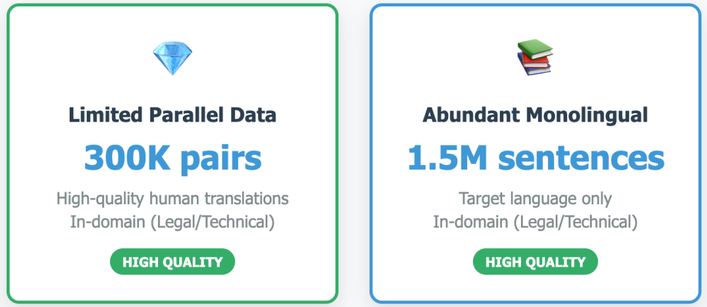
Target language only
In-domain (Legal/Technical)

HIGH QUALITY

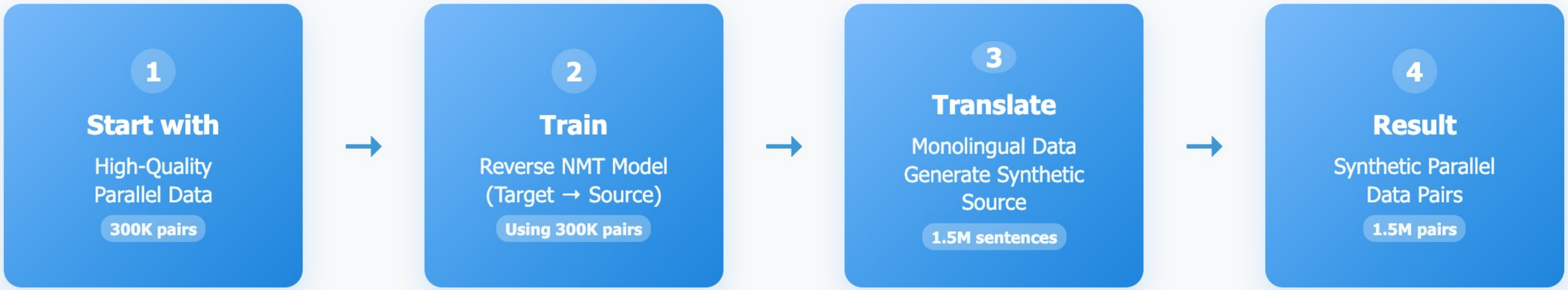
GHENT UNIVERSITY

Background: Back-Translation

📊 Data Sources in Specialized Domains

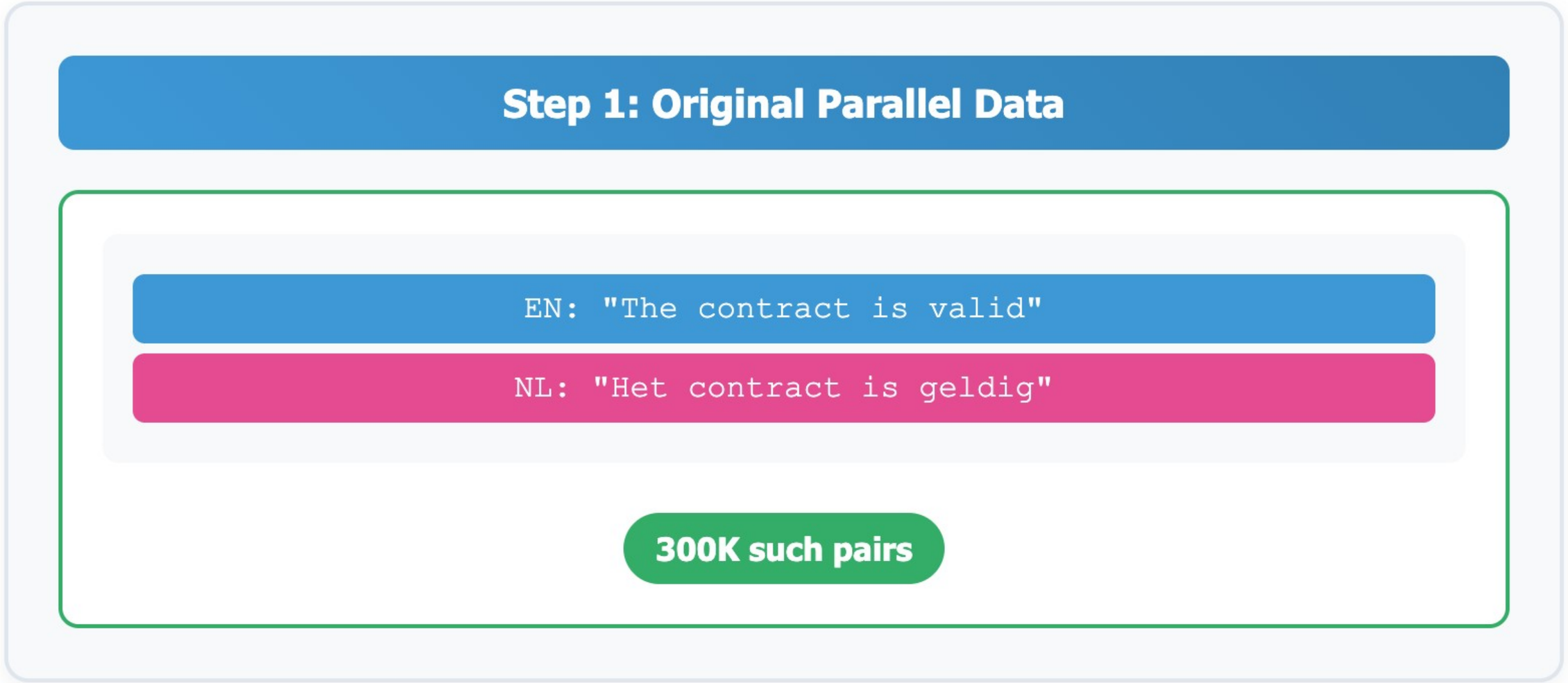


🔄 Back-Translation Process



GHENT UNIVERSITY

Background: Back-Translation



GHENT UNIVERSITY

Background: Back-Translation

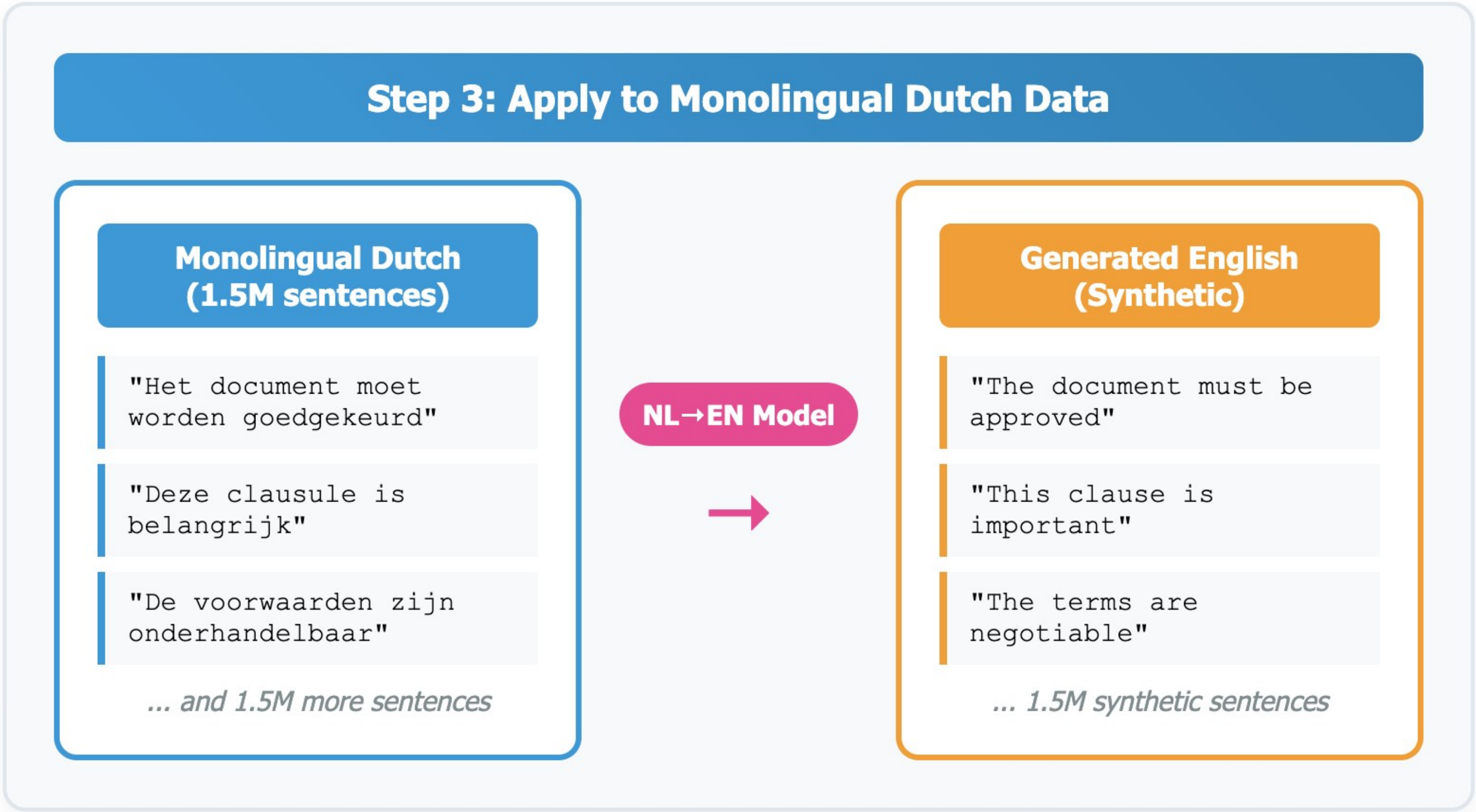
Step 2: Train Reverse Model (NL→EN)

Reverse NMT Model
NL → EN

Trained on 300K parallel pairs

GHENT UNIVERSITY

Background: Back-Translation



Background: Back-Translation

Step 4: Resulting Synthetic Parallel Data

New Synthetic Training Pairs

```
EN: "The document must be approved" (synthetic)
NL: "Het document moet worden goedgekeurd" (original)
```

```
EN: "This clause is important" (synthetic)
NL: "Deze clausule is belangrijk" (original)
```

1.5M new synthetic parallel pairs

GHENT UNIVERSITY

Background: Back-Translation

300K

Parallel Data

High Quality

+

1.5M

Back-Translated

Synthetic/Noisy

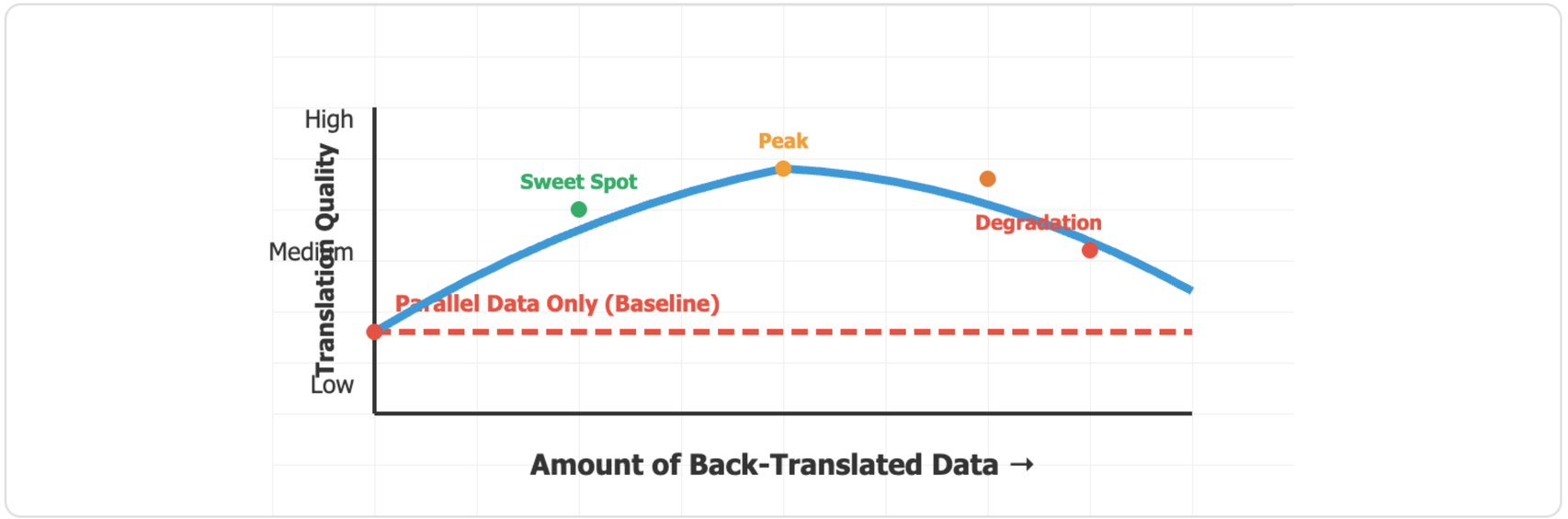
=

1.8M

Total Training

Mixed Quality

📈 Performance vs. Back-Translation Data Amount



Optimal Range:
- Back-translation significantly improves translation quality

Degradation Point:
- Too much synthetic data hurts performance due to noise

GHENT UNIVERSITY

Background: Fuzzy Matches

🔍 What are Fuzzy Matches?

Fuzzy Matches (FMs) are similar translations retrieved from translation memory or training data that can guide machine translation systems towards better outputs.

Example: Finding Fuzzy Matches

New Source:
"The contract is invalid"

85% similar

Fuzzy Match:
"Het contract is ongeldig"

New Source:
"The document must be signed"

78% similar

Fuzzy Match:
"Het document moet worden ondertekend"

Background: Fuzzy Matches

Architecture Modification

Modify the MT model architecture
extra components like attention la
special decoders, or FM-specific m

Cao and Xiong, 2018; He et al., 2020

Shared Goal

Steer MT output towards translations of fuzzy matches

All three approaches aim to leverage similar existing translations to improve the quality and consistency of machine translation systems.

Moslem et al., 2023

ata Augmentation IFR)

y Repair: Augment training data
nating source sentences with
translations. No architecture
eded.

<sep> fuzzy_target

ezcan, 2019

During Training

Step 1: Original Sentence Pair

Source (S): We found three studies for inclusion in the review.

Target (T): We vonden drie studies voor opname in de review.

Step 2: Retrieved Fuzzy Match (Similarity: 0.9309)

FMS: We identified **nine** eligible studies for inclusion in the review.

FMT: We identificeerden **negen geschikte** studies voor opname in de review.

Differences highlighted in bold

Step 3: Creating Augmented Input (S*)

Formula: S* = Source <sep> FMT

Actual S*:

We found three studies for inclusion in the review.
<sep> We identificeerden negen geschikte studies voor opname in de review.

Same process applied at inference time - source is augmented with FMT before translation

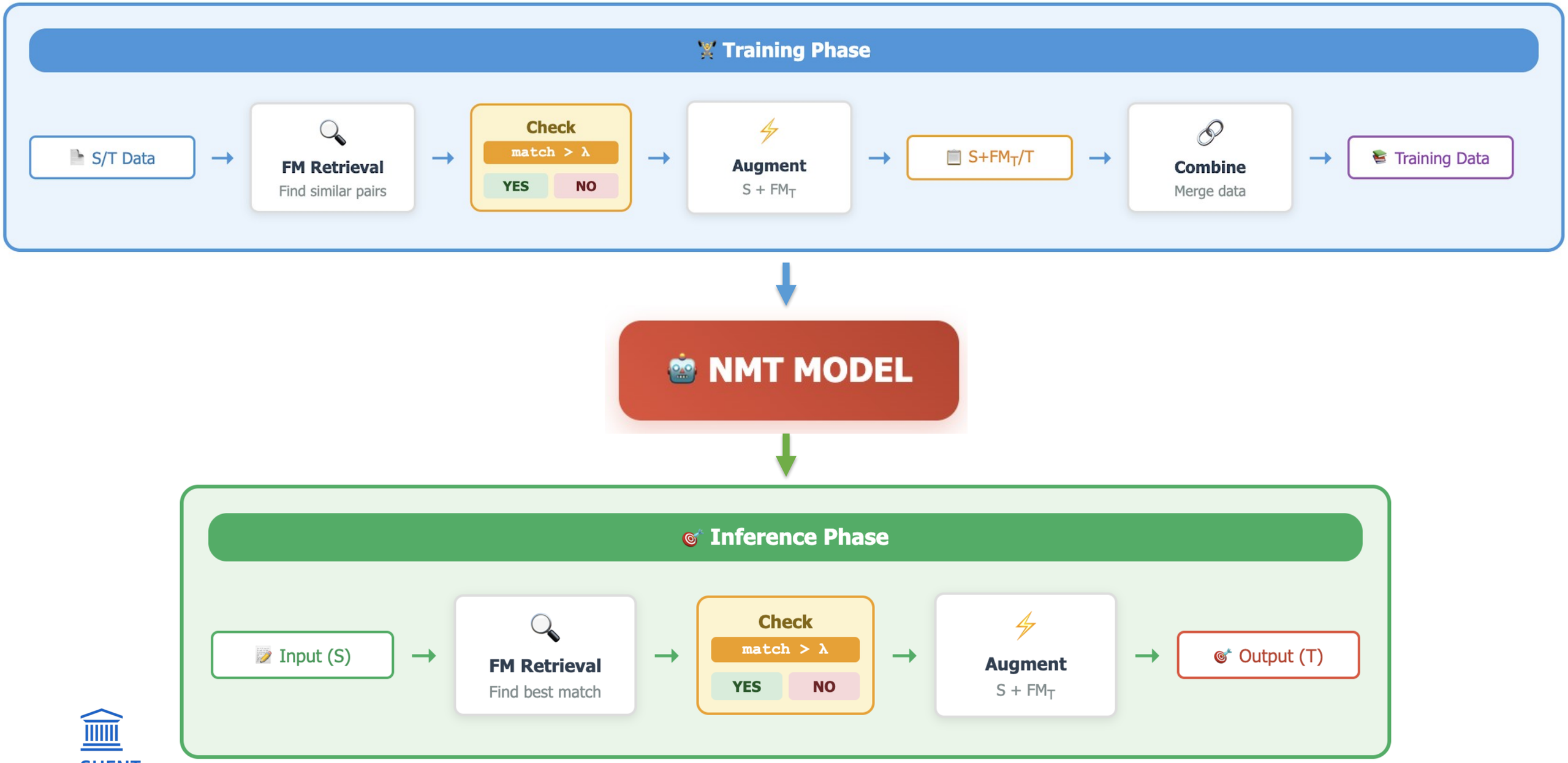
🔍 Fuzzy Matches Retrieved From

📚 Training Data

OR

💾 Translation Memory

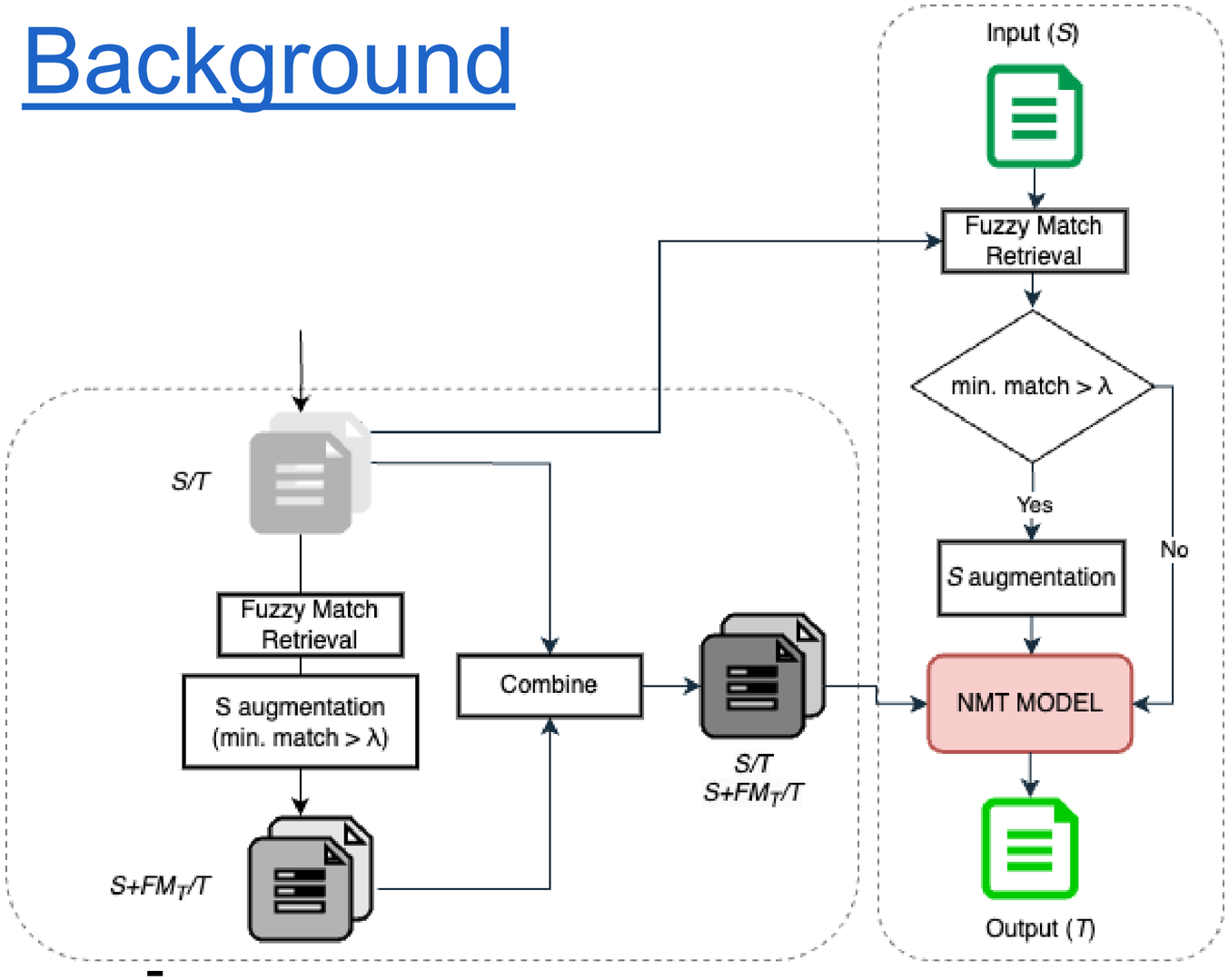
GHENT UNIVERSITY



GHENT UNIVERSITY

Background

Towards a better integration of fuzzy matches in neural machine translation through data augmentation [Bulté & Tezcan, 2019]



S	Debt, breakdown by residual maturity
$score$	0.9812
FM_S	Debt, breakdown by **initial** maturity
FM_T	Dette, ventilation par échéance **initiale**
S'	Debt, breakdown by residual maturity $<$ sep $>$ Dette, ventilation par échéance initiale
T	Dette, ventilation par échéance résiduelle

GHENT UNIVERSITY

Background

- FM augmentation results in optimal NMT performance in **high-resource scenarios** = (very) large bilingual training data [Bulté & Tezcan, 2019; Xu et al., 2023; Reheman et al., 2023]

- To remedy this limitation, **additional monolingual data in the target language** was used to retrieve FMs directly from texts in target language (via multilingual sentence embeddings) [Tamura et al., 2023]

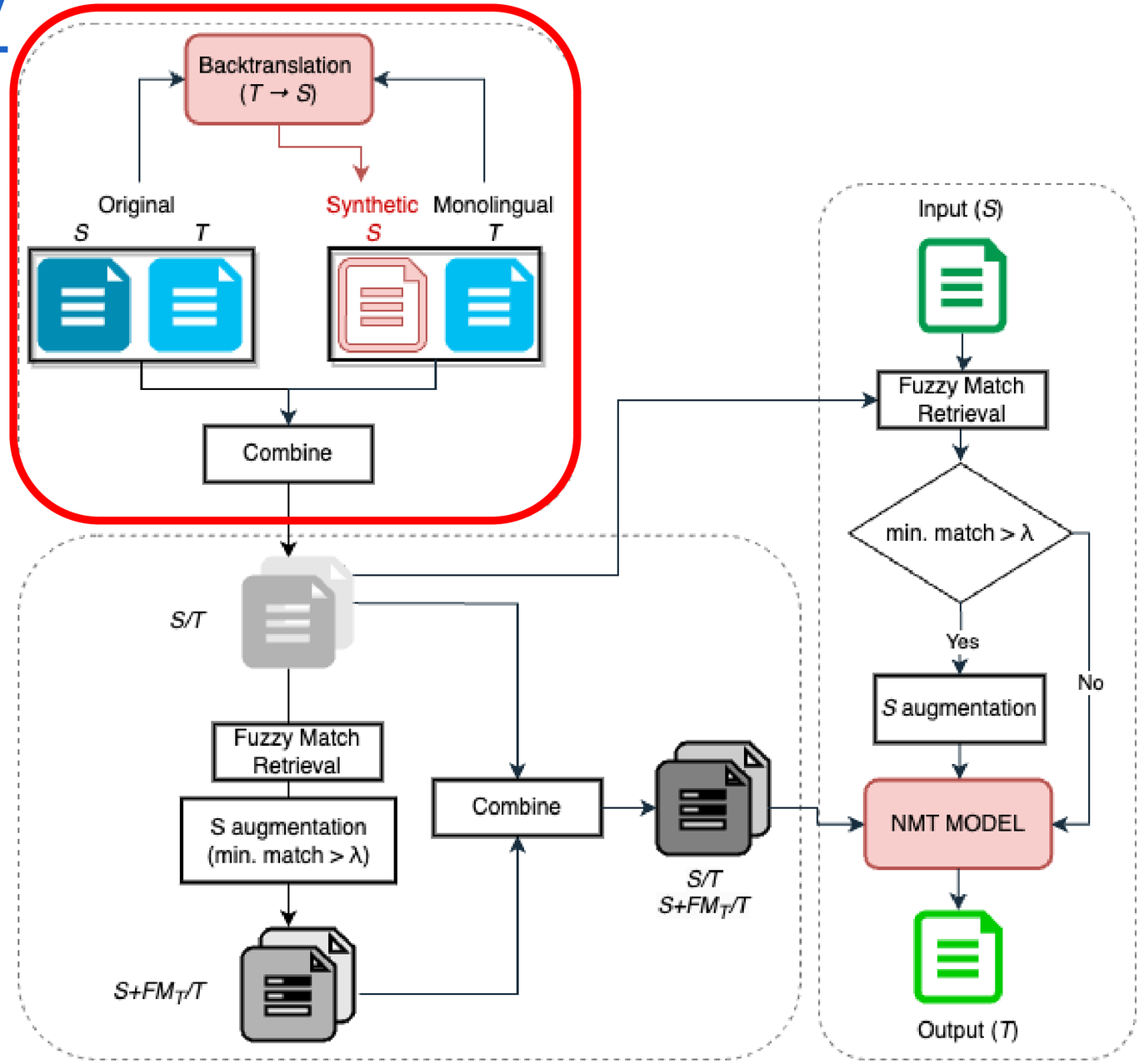
- Our goal is similar:
  - Extend the effectiveness of FM-augmentation **to lower-resource scenarios**
  - Focus = limited bilingual data, and **additional monolingual data**

- Our approach is different:
  1. MT additional monolingual data to source language (i.e. **back-translation**)
  2. Use the **synthetic bilingual data** as:
     - additional training data
     - for FM augmentation

GHENT UNIVERSITY

Proposed Method: Combining Back-Translation & Neural Fuzzy Repair

Methodology



B BASE

BASELINE

Standard NMT training using only original bilingual data.

Training Data:

- Original Bilingual Data ~300K pairs

BT BT

BASELINE

Back-translation approach: Original data plus synthetic bilingual data for training.

Training Data:

- Original Bilingual Data ~300K pairs
- Back-translated Data ~1.5M pairs

Total Training:

~1.8M sentence pairs

BT+NFR BT + NFR

PROPOSED

Our Approach: Original + additional synthetic bilingual data used for both training AND fuzzy match augmentation.

Training Data:

- Original Bilingual Data ~300K pairs
- Back-translated Data ~1.5M pairs
- + FM Augmentation

Key Innovation:

Synthetic data for Training + FM Pool

Total: ~1.8M pairs with FM augmentation

NFR NFR

BASELINE

Neural Fuzzy Repair: Original data with fuzzy match augmentation.

Training Data:

- Original Bilingual Data ~300K pairs
- + FM Augmentation

Augmentation:

Training + Inference

NFR NFRmono

BASELINE

NFR with additional monolingual data for FM retrieval during inference only.

Training Data:

- Original Bilingual Data ~300K pairs

Inference Only:

- + Monolingual Target Data
- + FM Augmentation

Augmentation:

Inference Only

GHENT UNIVERSITY

Methodology

Feature Comparison				
System	Original Data	Synthetic/HQ Data	FM Augmentation	Key Feature
BASE	✓	✗	✗	Standard NMT baseline
BT	✓	✓ (Synthetic)	✗	Back-translation only
NFR	✓	✗	✓	FM augmentation only
NFRmono	✓	✗	✓ (inference)	FM with monolingual data
BT+NFR (Ours)	✓	✓ (Synthetic)	✓	**Combined approach**
BASE_HQ	✓	✓ (High-Quality)	✗	**Oracle baseline**
NFR_HQ	✓	✓ (High-Quality)	✓	**Oracle upper bound**

GHENT UNIVERSITY

Methodology

🛠️ Training Framework

OpenNMT: Train all NMT models from scratch (no LLMs)

Evaluation Metrics

BLEU

Precision-based metric measuring n-gram overlap between hypothesis and reference translations

chrF

Character-level F-score that considers both precision and recall of character n-grams

COMET

Neural metric trained to predict human judgments of translation quality

Statistical Significance Testing

Bootstrap resampling to ensure reliable performance comparisons

GHENT UNIVERSITY

Data

DATA

English → Ukrainian

Legal Domain, Parliament Proceedings

Sources:

- EU acts • Ukrainian law
- Ukrainian law docs • Legal corpus

	Train	Val	Test
Bilingual	**286K**	2K	2K
Monolingual	**1.46M**	—	—

English ↔ French

European Legislation

Sources:

- DGT-TM (Bilingual)
- DGT-TM (Monolingual FR + EN)

	Train	Val	Test
Bilingual	**300K**	2K	2K
Mono (FR)	**1.5M**	—	—
Mono (EN)	**1.5M**	—	—

⚠️ NFR Limitation

~300K pairs

NFR not useful in low-resource setting

[Bulté & Tezcan, 2019]

✅ BT Benefits

~1.5M sentences

Up to 5x bilingual data improves results

[Edunov et al., 2018]

🎯 Lower-Resource

Test **reduced data sizes** to simulate even more constrained scenarios

DATA

Monolingual Data Reduction

Vary the amount of monolingual data while keeping bilingual data constant at ~300K pairs.

Monolingual Data Amounts:

100%	1.5M sentences
80%	1.2M sentences
60%	900K sentences
40%	600K sentences
20%	300K sentences

Purpose:

Test how much monolingual data is needed for effective back-translation in FM augmentation.

Bilingual Data Reduction

Vary the amount of bilingual data while keeping monolingual data constant at ~1.5M sentences.

Bilingual Data Amounts:

100%	300K pairs
66%	200K pairs
50%	150K pairs
33%	100K pairs

Purpose:

Test robustness in extremely low-resource scenarios where even less bilingual data is available.

Results & Conclusion

Results

1. Using all available data (bilingual & monolingual)

System	EN→UK			EN→FR			FR→EN		
	BLEU	chrF	COMET	BLEU	chrF	COMET	BLEU	chrF	COMET
BASE	54.90	75.68	90.79	45.76	64.97	79.75	47.76	65.19	80.69
BT (Sennrich et al., 2016)	56.18	76.63	91.61	49.12	67.19	81.42	50.42	67.68	82.32
NFR (Tezcan et al., 2021)	57.71	77.23	91.34	45.61	64.91	79.90	48.05	65.44	80.82
NFR$_{mono}$ (Tamura et al., 2023)	60.26	78.58	91.55	46.73	65.48	79.88	49.02	66.05	81.05
Proposed system	**66.79**	**81.99**	**92.23**	**57.07**	**71.83**	**82.82**	**60.04**	**72.96**	**83.55**

Baseline Analysis

- ✓ **BT** is the best baseline for EN ↔ FR
- ✓ **NFRmono** is the best baseline for EN → UK
- ✓ **NFR** not useful for EN ↔ FR (confirmed)
- ✓ **NFR** useful for EN → UK

Our System Performance

- ✓ Outperforms **all baseline systems**
- ✓ Up to **+9.62 BLEU** over best baseline
- ✓ Up to **+12.28 BLEU** over BASE
- ✓ Consistent improvements across all language pairs

GHENT UNIVERSITY

Results

1. Using all available data (bilingual & monolingual)

System	EN→UK			EN→FR			FR→EN		
	BLEU	chrF	COMET	BLEU	chrF	COMET	BLEU	chrF	COMET
BASE	54.90	75.68	90.79	45.76	64.97	79.75	47.76	65.19	80.69
BT (Sennrich et al., 2016)	56.18	76.63	91.61	49.12	67.19	81.42	50.42	67.68	82.32
NFR (Tezcan et al., 2021)	57.71	77.23	91.34	45.61	64.91	79.90	48.05	65.44	80.82
NFR_{mono} (Tamura et al., 2023)	60.26	78.58	91.55	46.73	65.48	79.88	49.02	66.05	81.05
Proposed system	**66.79**	**81.99**	**92.23**	**57.07**	**71.83**	**82.82**	**60.04**	**72.96**	**83.55**
BASE_HQ	–	–	–	53.94	70.27	83.18	54.90	70.38	83.87
NFR_HQ	–	–	–	57.75	72.57	83.91	60.63	73.68	84.48

🏆 NFR_HQ Performance

NFR_HQ performs best (as expected for high-resource scenarios)

Our proposed system performs **very close** to NFR_HQ:

- EN→FR: 57.07 vs 57.75 (-0.68 BLEU)
- FR→EN: 60.04 vs 60.63 (-0.59 BLEU)

Using synthetic data instead of high-quality human translations does not drastically impair translation quality.

📈 BASE_HQ Comparison

Our proposed system **outperforms BASE_HQ** in BLEU/chrF:

- EN→FR: 57.07 vs 53.94 (+3.13 BLEU)
- FR→EN: 60.04 vs 54.90 (+5.14 BLEU)

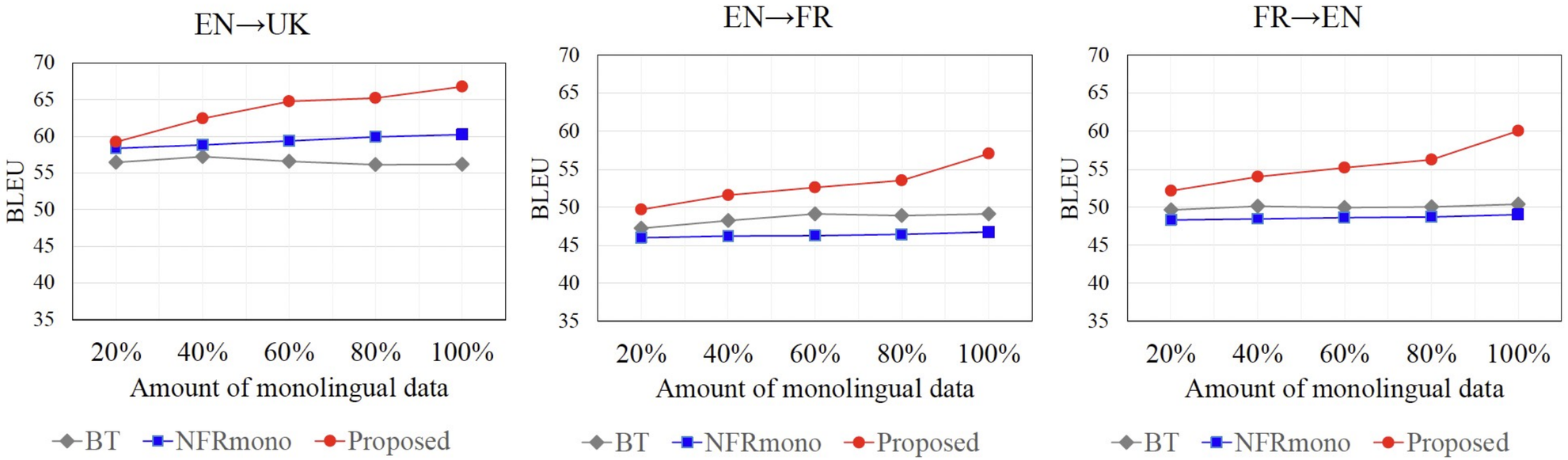
Note: BASE_HQ performs better on COMET

We can build better NMT systems with synthetic data + FM augmentation compared to vanilla NMT.

GHENT UNIVERSITY

Results

2. Using reduced monolingual data sets (keeping all bilingual data)



System Behavior Trends

- **BT** reaches a plateau around 40-60% of data size
- **NFRmono** keeps slightly improving with additional data
- Both baselines show diminishing returns with more data
- Clear performance differences emerge early

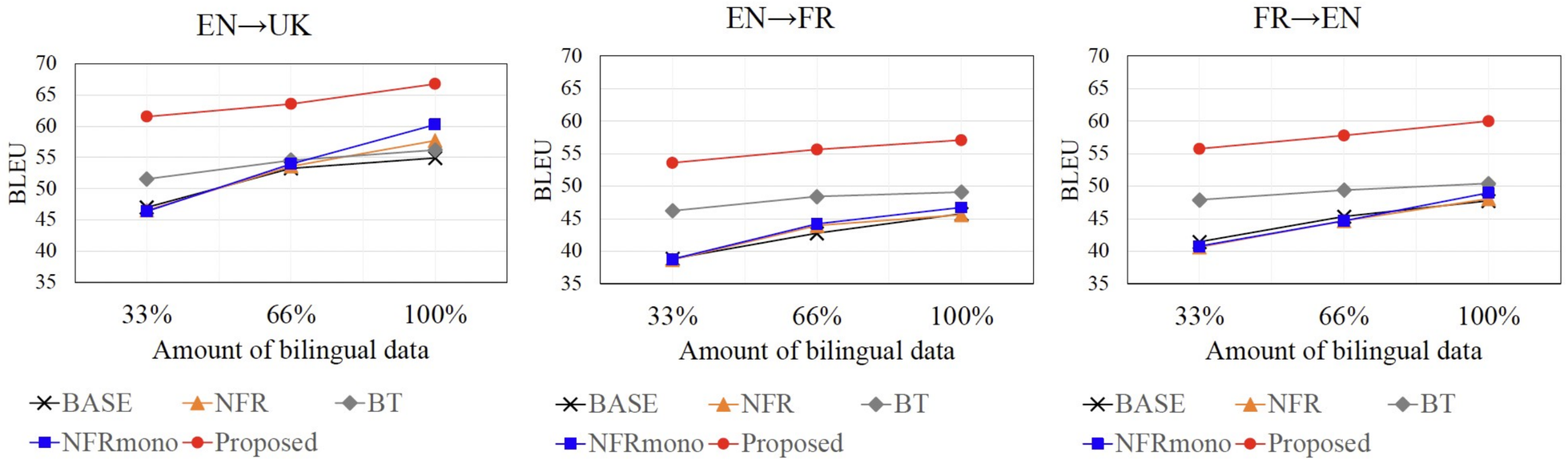
Our System Performance

- **Outperforms both baselines** in all data conditions
- **Largest improvements** with maximum available data
- **Consistent upward trend** across all language pairs
- **Robust performance** even with limited monolingual data

GHENT UNIVERSITY

Results

3. Using reduced bilingual data sets (keeping all monolingual data)



📈 **All baselines improve** with larger bilingual data, confirming expected scaling behavior

🎯 **NFRmono benefits most** from larger bilingual data among baseline methods

🚀 **Proposed system outperforms** all baselines across all data scenarios

💪 **Large relative improvements** maintained even in lowest resource scenario (33% data)

GHENT UNIVERSITY

Conclusions

- A simple, yet effective approach for improving NMT systems
- Backtranslation + FM-augmentation
  - No changes required to the underlying NMT architecture (data augmentation)
  - Wider adoption than alternatives?
- Additional monolingual data sets in the target language can be more efficiently than existing alternatives
- Large (statistically significant) improvements compared to best alternative in all data configurations (including lower resource scenarios)
- Using high-quality, large data sets instead does not improve the results drastically
- Similar trends observed across different metrics

GHENT UNIVERSITY

Limitations & Future Work

Limitations:

- Only 2-domains and 3 language pairs tested
- Only automated evaluation (no human evaluation)

Future work:

- Can similar improvements be observed using FM-augmentation in LLMs?
  - Using LLMs for back-translating monolingual data
  - Integrating FMs through in-context learning
- Can we generate additional monolingual data in the target language, when they are not available?
  - Use LLMs for domain-specific data generation
- Can we even use this approach when only monolingual data sets are available in the target language (no bilingual data)?

GHENT UNIVERSITY

REFERENCES

- Qian Cao and Deyi Xiong. 2018. Encoding gated translation memory into neural machine translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 646 3042–3047.
- Qiuxiang He, Guoping Huang, Qu Cui, Li Li, and Lemao Liu. 2021. Fast and accurate neural machine translation with translation memory. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3170–3180, Online. Association for Computational Linguistics.
- Maxime Bouthors, Josep Crego, and François Yvon. 2023. Towards example-based NMT with multi-Levenshtein transformers. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 1830–1846, Singapore. Association for Computational Linguistics.
- Bram Bulté and Arda Tezcan. 2019. Neural fuzzy repair: Integrating fuzzy matches into neural machine translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1800–1809, Florence, Italy. Association for Computational Linguistics.
- Jitao Xu, Josep Crego, and Jean Senellart. 2020. Boosting neural machine translation with similar translations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1580–1590, Online. Association for Computational Linguistics.
- Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

FACULTY OF ARTS AND PHILOSOPHY

Thomas Moerman

PhD Researcher

LANGUAGE AND TRANSLATION TECHNOLOGY TEAM (LT3)

✉ Thomas.Moerman@UGent.be

Tezcan, A., Skidanova, A., & Moerman, T. (2024). Improving Fuzzy Match Augmented Neural Machine Translation in Specialised Domains through Synthetic Data. *The Prague Bulletin of Mathematical Linguistics*, *122*, 9–42.

www.ugent.be

https://lt3.ugent.be/

GHENT UNIVERSITY

language and translation technology team