Leveraging Synthetic Monolingual Data for Fuzzy-Match Augmentation in Neural Machine Translation: A Preliminary Study

Thomas Moerman and **Arda Tezcan**
Language and Translation Technology Team (LT^3)
Ghent University, Belgium
{thomas.moerman, arda.tezcan}@ugent.be

1 Background and Methodology

Recent work has demonstrated that specialized neural machine translation (NMT) models, as well as Large Language Models (LLMs), can utilize fuzzy matches (FMs) (i.e., similar translations for a given source sentence) effectively to produce translations of higher quality (Xu et al., 2020; Tezcan et al., 2021; Moslem et al., 2023).

Earlier studies have shown that FM-augmentation is especially useful in domain-specific scenarios where large bilingual datasets are available (Bulté and Tezcan, 2019; Xu et al., 2020). A more recent study (Tezcan et al., Under Revision) further demonstrated the effectiveness of FM-augmentation in settings where this approach alone is not helpful due to the availability of limited bilingual data sets by using additional monolingual data available in the target language through back-translation (BT) (Sennrich et al., 2015; Edunov et al., 2018) and subsequently applying the Neural Fuzzy Repair (NFR) technique for FM-augmentation, which relies on concatenating source sentences with the translations FMs (Tezcan et al., 2021).

This study further investigates the usefulness of FM-augmentation for NMT in domain-specific scenarios where limited bilingual datasets are available without any additional monolingual datasets. We aim to bridge this gap by generating additional monolingual data in the target language using an LLM and employing back-translation to generate corresponding sentences in the source text, as also proposed by Moslem et al. (2022). Additionally, we use the synthetic source/target sentence pairs for FM-augmentation in the context of specialized NMT systems.

In this preliminary study, we use the DGT Translation Memory (DGT-TM) of the European Commission's translation service[1], for English→French, covering European legislation texts. The dataset includes 300,000 sentence pairs for NMT training and 2,000 for validation and testing. The choice of this data set is two-fold: (i) it has been demonstrated that the NFR approach itself did not yield performance improvements when using this data set size obtained from the DGT-TM (Bulté and Tezcan, 2019), and (ii) the NFR approach yielded clear improvements when the training data was increased through back-translating the additionally available (high-quality) monolingual data in the target language (Tezcan et al., Under Revision).

The proposed approach consists of three main steps:

1. **Synthetic Data Generation:** First, synthetic sentences in the target language (French) are produced using the Mistral-7b-instruct-v0.2 model (Jiang et al., 2023), following a prompt designed to achieve thematic coherence (Veselovsky et al., 2023). This stage employs the vLLM library[2], which utilizes paged attention (Kwon et al., 2023). Further details on the synthetic data generation process are provided in Appendix A.1.

2. **Back-translation:** Next, these synthetic sentences are back-translated into the source language using a pre-trained NMT system with the same training data (300K sentence pairs), only trained in the reverse language direction (FR→EN). The synthetically generated bilin-

[1] https://opus.nlpl.eu/DGT/corpus/version/DGT
[2] https://github.com/vllm-project/vllm

Proceedings of the First International Workshop on Knowledge-Enhanced Machine Translation, pages 34–39
June 27, 2024 ©2024 European Association for Machine Translation

gual data set is then merged with the original training data.

3. **FM Augmentation:** This step involves using the NFR approach (Tezcan et al., 2021), which retrieves the highest FM for each source sentence from the merged training data and uses its translation for source-augmentation in each data partition, where FM similarity is measured by cosine similarity between sentence embeddings[3].

To test the usefulness of the proposed approach in different data settings, the training data was incrementally increased through synthetic data generation in the target language from 300K sentences (the same size as the bilingual data set) to 1.5M sentences (five times larger than the bilingual data set size).

We primarily utilized the default settings of the transformer architecture as implemented in OpenNMT[4] (Klein et al., 2017) with early stopping. SacreBLEU (Post, 2018), ChrF (Popović, 2015) and COMET (Rei et al., 2020) were used to automatically assess the MT performance.

2 Preliminary Results

The preliminary results of this ongoing study highlight several key findings:

- Applying FM-augmentation (NFR) on the original bilingual training data does not yield better translation performance against the standard (baseline) NMT system, confirming previous findings (Bulté and Tezcan, 2019).

- Utilizing additional synthetic training data without FM-augmentation, namely synthetically generated monolingual data in the target language via Mistral and corresponding source sentences produced through BT, achieves results comparable to the baseline NMT system.

- Using FM-augmentation in combination with synthetic data generation improves results across all additional monolingual data set sizes, outperforming both the baseline and NFR systems.

- The proposed approach achieves optimal improvements when the synthetically generated monolingual data set size is twice (BLEU and ChrF) or four times (COMET) that of the original bilingual data set. However, performance declines with the addition of larger synthetic data sets.

- The optimal improvements when using the proposed approach are observed to be up to +1.44 BLEU points compared to the baseline NMT system and +1.59 BLEU points compared to the NFR system while showing statistically significant improvements across all three metrics (bootstrap resampling with $p < 0.05$).

For an overview of the automated evaluation results for each system tested in this study, please see Appendix A.3.

Preliminary results from this ongoing work suggest that in this specific setting, the proposed approach, consisting of generating (i) synthetic monolingual data in the target language via an LLM, (ii) synthetic source sentences through back-translation, and (iii) applying NFR, could be an effective strategy for enhancing the performance of specialized NMT systems.

The effectiveness of the proposed approach prompts further investigation into whether (i) similar observations can be made in different data settings (especially in lower-resource settings), domains and language directions; and (ii) the MT performance can be further enhanced through alternative synthetic data generation strategies (both in the target and source language) and/or with increasing amounts of such additional synthetic data.

Acknowledgements

The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), which is funded by Ghent University, FWO and the Flemish Government department EWI.

References

Bulté, Bram 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, July. Association for Computational Linguistics.

Edunov, Sergey, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at

[3]https://github.com/lt3/nfr

[4]https://github.com/OpenNMT/OpenNMT-py-v3.5.1

scale. In Riloff, Ellen, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium, October-November. Association for Computational Linguistics.

Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. *Computing Research Repository*, arXiv:1701.02810.

Kwon, Woosuk, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Moslem, Yasmin, Rejwanul Haque, John Kelleher, and Andy Way. 2022. Domain-specific text generation for machine translation. In Duh, Kevin and Francisco Guzmán, editors, *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 14–30, Orlando, USA, September. Association for Machine Translation in the Americas.

Moslem, Yasmin, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In Nurminen, Mary, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland, June. European Association for Machine Translation.

Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. In Bojar, Ondřej, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October. Association for Computational Linguistics.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Tezcan, Arda, Bram Bulté, and Bram Vanroy. 2021. Towards a better integration of fuzzy matches in neural machine translation through data augmentation. *Informatics*, 8(1).

Tezcan, Arda, Alina Skidanova, and Thomas Moerman. Under Revision. Improving fuzzy match augmented neural machine translation through synthetic data.

Veselovsky, Veniamin, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. 2023. Generating faithful synthetic data with large language models: A case study in computational social science.

Xu, Jitao, Josep Crego, and Jean Senellart. 2020. Boosting neural machine translation with similar translations. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1580–1590, Online, July. Association for Computational Linguistics.

A Appendix

A.1 Synthetic Data Generation

Sampling Parameters for Mistral-7b-instruct-v0.2

The specific sampling parameters differing from default values are outlined below. For details on default parameter settings, please refer to the vLLM library documentation at https://docs.vllm.ai/en/latest/dev/sampling_params.html. These parameter adjustments were adopted from the findings in Moslem et al. (2022).

Parameter	Value
Top-p	0.95
Top-k	50
Frequency Penalty	0.5
Repetition Penalty	1.2
Max Tokens	400

Table 1: Sampling parameters for Mistral-7b-instruct-v0.2

Prompt Design

Table 2 outlines the specific prompt design utilized for generating French sentences, highlighting the instruction and the examples given to the language model and the response given to that prompt.

A.2 FM-augmentation

See Table 3 for an example of FM retrieval and source augmentation.

A.3 Translation Performance

See Table 4 for all metrics (BLEU, ChrF and COMET) and Table 5 for the performance of the back-translation model.

- **Baseline** refers to the standard NMT system trained on the original 300k bilingual data.
- **Baseline + BT(1:X)** refers to the non-augmented NMT system, using additional synthetically generated target sentences (through LLM) and their translations in the source language (through back-translation), where X indicates the ratio of synthetic to original data.
- **NFR** refers to the system trained on the original bilingual data after applying FM-augmentation.
- **BT 1:X + NFR** refers to the FM-augmented NMT model, using additional synthetically generated target sentences (through LLM) and their translations in the source language (through back-translation) with Neural Fuzzy Repair, where X indicates the ratio of synthetic to original data (the proposed approach in this study).

Prompt (French)	LLM Response
[INST] Votre tâche consiste à générer des phrases françaises similaires en termes de domaine et de style aux exemples fournis ci-dessous. Exemples de phrases : 1. Les cotisations des membres pour l'exercice au cours duquel ils deviennent membres du Conseil oléicole international sont exigibles à la date où ils deviennent membres. 2. Au vu de ces circonstances, la Suède a été confrontée à des difficultés temporaires d'approvisionnement général en semences de plantes fourragères de printemps et en semences de céréales de printemps, et ces difficultés ne devraient pas se dissiper. 3. La contribution de chaque État membre participant auprès duquel une contribution est demandée est proportionnelle à la part du RNB de cet État membre dans le total des RNB des États membres participants auprès desquels une contribution est demandée. 4. L'article 199 ter est remplacé par le texte suivant: 5. Dans les limites du montant fixé, l'organe de la société habilité à cet effet décide, le cas échéant, d'augmenter le capital souscrit. Veuillez générer 10 nouvelles phrases françaises basées sur les exemples ci-dessus. [/INST]	1. Les impôts des nouveaux membres du Comité céréalier international sont dus à la date où ils rejoignent ce comité. 2. En raison de ces circonstances, la Norvège a confronté des difficultés temporaires d'approvisionnement général en graines de fourrages d'été et en graines de blé d'été, et ces difficultés ne devraient pas disparaître. 3. La contribution de chaque État membre participant, où une contribution est demandée, est proportionnelle à la part du Fonds communautaire de ce pays membre dans le total des Fonds communautaires des États membres participants where a contribution is requested. 4. L'article 200 quater est remplacé par le texte suivant : 5. Dans les limites du montant fixé, l'organe de la société autorisé à cette fin décide, le cas échéant, d'augmenter le capital souscrit. 6. Les cotisations des nouveaux membres du Conseil oléicole européen sont exigibles à la date où ils deviennent membres. 7. En raison des circonstances actuelles, la Finlande a été confrontée à des difficultés temporaires d'approvisionnement général en graines de fourrages hiver et en graines de blé d'hiver, et ces difficultés ne devraient pas se résoudre. 8. La contribution de chaque État membre participant où une contribution est demandée est proportionnelle à la part du Fonds communautaire de ce pays membre dans le total des Fonds communautaires des États membres participants where a contribution is demanded. 9. L'article 201 quinquiès est modifié pour devenir le texte ci-après : 10. Dans les limites du montant prévu, l'organe de la société compétent en matière de cession de droits décide, si nécessaire, d'augmenter le capital souscrit.

Table 2: Prompt used for generating domain-specific French sentences alongside the LLM-generated response.

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

Table 3: An example of FM retrieval and source augmentation (S') for a given source sentence (S) for the EN→FR language direction, with the translation 'T'. 'FM_S' and 'FM_T' refer to the source and target sides of the retrieved FM, respectively. The sentence similarity score is indicated as '$score$'.

Configuration	**BLEU**	**ChrF**	**COMET**
Baseline	45.76	64.97	79.75
BT 1:1	45.69	65.11	80.20
BT 1:2	45.79	65.26	80.44
BT 1:3	44.96	64.70	80.43
BT 1:4	44.57	64.44	80.31
BT 1:5	45.19	64.89	80.64
NFR	45.61	64.91	79.90
BT 1:1 + NFR	47.14	65.91	80.76
BT 1:2 + NFR	**47.20**	**66.03**	80.76
BT 1:3 + NFR	47.03	65.90	80.90
BT 1:4 + NFR	46.87	65.80	**80.91**
BT 1:5 + NFR	45.90	65.52	80.88

Table 4: Automated evaluation of the different NMT systems.

System	**BLEU**	**ChrF**	**COMET**
FR→EN	47.76	65.19	80.69

Table 5: Automated evaluation of the back-translation (NMT) model, which is trained on the original parallel data set in reverse language direction and evaluated on the reversed test set.