UDC 004.85:004.912

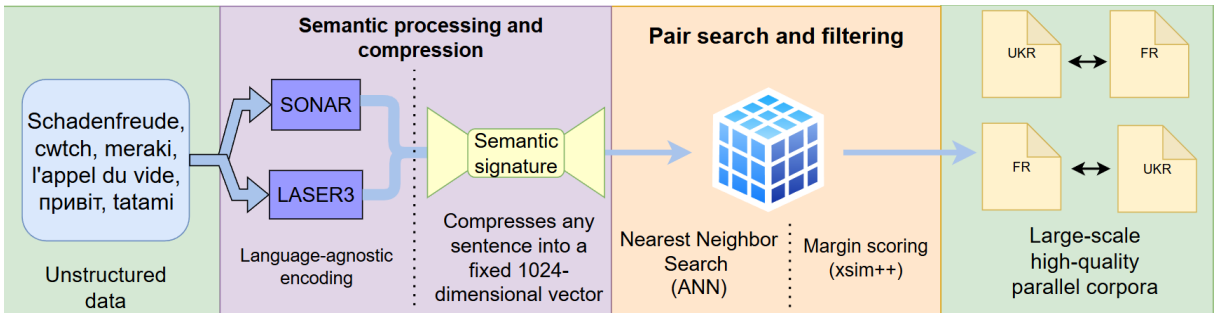
METHODS FOR CREATING DATASETS FOR SMALL-RESOURCE LANGUAGE PAIRS

Lysenko D.O.
e-mail: danylo.lysenko@nure.ua
Research Advisor – PhD, senior lecturer Kobylin I.O.
Kharkiv National University of Radio Electronics, Dept. of ITM
Kharkiv, Ukraine

This research addresses the critical challenge of data scarcity in developing direct neural machine translation (NMT) systems for low-resource language pairs without relying on a high-resource pivot language like English. To overcome extreme data limitations, the research proposes a fully automated pipeline that combines bitext mining via language-agnostic embedding spaces, synthetic data generation using Large Language Models (LLMs), and internal structural data augmentation. By integrating these generative methodologies with rigorous automated noise filtering, the proposed approach successfully scales microscopic seed dictionaries into robust, high-quality parallel corpora suitable for training accurate and culturally authentic translation models.

Historically, machine translation has avoided low-resource languages, and the use of English as an intermediary has led to the accumulation of errors and loss of context. The solution is to create massive datasets for direct translation using fully automated pipeline-type systems. The combination of semantic search, artificial intelligence generation, and internal augmentation allows for exponential growth in data volume even under conditions of extreme resource scarcity.

Automated bi-text mining extracts data from monolingual web corpora (e.g., Wikipedia or Common Crawl) using multilingual embedding spaces such as SONAR or LASER3. These models convert sentences from different languages into universal vectors of the same size, allowing semantically identical phrases to be automatically combined into ready-made parallel pairs by searching for the closest mathematical neighbours.



Picture 1 – Visualisation of the bi-text mining process

When there are no comparable texts on the internet, large language models (LLMs) are capable of generating synthetic parallel data directly between two

target languages. This direct transfer significantly reduces the structural biases commonly introduced by pivoting through a high-resource intermediate language. To achieve this, they use fragment prompting (Fragment-Shot), which breaks down the input sentence and searches for exact syntactic matches in the base dictionary. By forcing the model to translate without switching to English, the system preserves cultural nuances and idiomatic expressions that might otherwise be lost in translation.

For autonomous improvement of the quality of generated synthetic data, iterative self-learning (SPIN) applies by minimizing the reliance on extensive human-annotated datasets. A weak base model generates an array of translations and constantly learns to distinguish its own errors from human benchmarks through a continuous feedback loop. As the model effectively acts as its own critic and refines its output distribution to match the target data, it independently improves the overall fluency and quality of texts with each new iteration.

By leveraging vast amounts of readily available unilingual data, the iterative back-translation and forward translation method effectively bridges the data gap for low-resource language pairs. This technique cyclically converts ordinary monolingual texts into pseudo-parallel texts using basic bidirectional models. During its optimization phase, the model relies on a tagging mechanism to correctly distinguish between authentic human language and machine-generated structures. Adding these special tags to generated sentences prevents the learning of artificial errors, allowing models to safely expand their dataset and become more accurate with each round of joint training.

When there is an extreme shortage of source texts, the MADLIBS algorithm performs internal augmentation of the existing micro-corpus to multiply the available training data without requiring any external text gathering. It creates a pseudo-dictionary by aligning words and dynamically replacing terms in sentences by parts of speech. This template-based substitution dramatically increases lexical diversity and generates new syntactic patterns. Through this process, the model gains significant robustness against rare vocabulary and unexpected grammatical constructions in real-world applications.

Without any parallel data, uncontrolled machine translation (UMT) linearly aligns the monolingual vector spaces of two different languages to enable seamless cross-lingual mapping even in the complete absence of paired sentence examples. To get rid of the inaccuracies of such literal translation, noise-cancelling autoencoders are used, which learn to grammatically reconstruct sentences from deliberately jumbled words. By mastering this denoising process, the system successfully learns the deep latent representations and structural rules of both languages independently.

Maintaining high corpus hygiene is crucial, as incorporating degraded or misaligned synthetic data can cause severe performance drops in neural networks. Because of this vulnerability to noise, all generated arrays undergo rigorous automated filtering using tools such as OpusFilter or Bicleaner. They remove

anomalies in length, foreign language inclusions, and check for cross-language similarity. This stringent quality control guarantees that the final model is trained exclusively on highly reliable and structurally sound parallel corpora by leaving only perfectly aligned texts for final training.

Table 1 – Summary of steps for creating a dataset

Data set creation stage	Methods and tools used
1. Basic data collection	Automated mining of bi-texts (SONAR, LASER3, LaBSE)
2. Initial generation	LLM prompting (Fragment-Shot), uncontrolled translation (UMT)
3. Scaling	Iterative back-translation, SPIN self-learning
4. Internal augmentation	Lexical substitution (MADLIBS), syntactic templates
5. Cleaning and curation	Heuristic rules, OpusFilter, Bicleaner, BLASER

Creating massive datasets for direct translation between low-resource languages is entirely achievable thanks to the pipeline synergy of semantic mining, LLM generation, and internal augmentation. Using these methods together with strict filtering completely eliminates the need for English as an intermediary language, allowing for the construction of high-quality and culturally authentic translation systems.

References:

1. SONAR (arXiv:2308.11466): Duquenne, P.-A., et al. (2023). SONAR: Sentence-level multimodal and language-agnostic representations. arXiv. https://doi.org/10.48550/arXiv.2308.11466

2. SPIN (arXiv:2401.01335): Chen, Z., Deng, Y., Yuan, H., Ji, K., & Gu, Q. (2024). Self-play fine-tuning converts weak language models to strong language models. arXiv. https://doi.org/10.48550/arXiv.2401.01335

3. Data Augmentation (ACL Anthology: P17-2090): Fadaee, M., Bisazza, A., & Monz, C. (2017). Data augmentation for low-resource neural machine translation. In R. Barzilay & M.-Y. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 567–573). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-2090