open access
Published online on: https://journal.uii.ac.id/jurnalsnati/
Check for updates
DOI: 10.20885/snati.v5.i2.47528
JURNAL SNATI Sains, Nalar, dan Aplikasi Teknologi Informasi
Jurnal Sains, Nalar, dan Aplikasi Teknologi Informasi
Vol. 5 No. 2 (2026) 138 - 145 ISSN Media Electronic: 2807-5935

Tinjauan Sistematis Transfer Learning dan Data Augmentasi pada Neural Machine Translation Bahasa Sumber Daya Rendah

A Systematic Review of Transfer Learning and Data Augmentation in Neural Machine Translation of Low-Resource Languages

Nur Fikri Khuluq[1], Muh Naufal Muzhaffar[2], Shofwatul 'Uyun[3*]
[1,2,3] Prodi Sarjana Informatika, Fakultas Sains dan Teknologi, UIN Sunan Kalijaga, Yogyakarta, Indonesia
[1]24206052001@student.uin-suka.ac.id, [2]24206052008@student.uin-suka.ac.id, [3*]shofwatul.uyun@uin-suka.ac.id

Abstract

Modern Neural Machine Translation (NMT) systems have achieved state-of-the-art, performance, largely due to the availability of large-scale parallel corpora. However, the translation quality of NMT for Low-Resource Languages (LRL) remains limited due to data sparsity. Numerous studies have proposed different strategies to address this challenge. Among the most widely adopted strategies are Transfer Learning (TL) and Data Augmentation (DA) strategies. This research aims to present a systematic review of how these techniques, including Back-Translation (BT), Hybrid Transfer Learning (HTL), and the utilization of self-supervised objectives such as Masked Language Modeling (MLM), Causal Language Modeling (CLM), and Denoising Autoencoder (DAE), affect the quality improvement of NMT for LRL. The findings show that a hybrid combination of TL and DA with a self-supervised objective is the most effective solution for extremely low-resource scenarios, capable of producing the highest translation quality (highest BLEU score) and outperforming baseline models and traditional methods such as Statistical Machine Translation (SMT).

Keywords: Transfer Learning; Data Augmentation; Neural Machine Translation; Self-Supervised Learning

Abstrak

Sistem Penerjemahan Mesin Neural (NMT) modern telah mencapai kinerja state-of-the-art, terutama ketika dilatih menggunakan korpora paralel berskala besar. Namun, kualitas terjemahan NMT untuk bahasa dengan sumber daya rendah (Low-Resource Languages/LRL) masih terbatas akibat kelangkaan data (data sparsity). Berbagai pendekatan telah dikembangkan untuk mengatasi tantangan ini, dengan Transfer Learning (TL) dan Data Augmentation (DA) menjadi dua strategi yang paling menonjol. Penelitian ini menyajikan tinjauan sistematis mengenai efektivitas teknik-teknik tersebut, termasuk Back-Translation (BT), Hybrid Transfer Learning (HTL), serta pemanfaatan objektif self-supervised seperti Masked Language Modeling (MLM), Causal Language Modeling (CLM), dan Denoising Autoencoder (DAE), dalam meningkatkan kualitas NMT untuk LRL. Temuan menunjukkan bahwa penggabungan hibrida antara TL dan DA dengan objektif self-supervised merupakan solusi paling efektif untuk skenario sumber daya yang sangat rendah (extremely low-resource), yang mampu menghasilkan kualitas terjemahan tertinggi (skor BLEU tertinggi) serta melampaui model dasar (baseline) maupun metode tradisional seperti Statistical Machine Translation (SMT).

Kata kunci: Transfer Learning; Data Augmentasi; Neural Machine Translation; Self-Supervised Learning

1. Pendahuluan

Penerjemahan Mesin (*Machine Translation* atau MT) merupakan teknologi kunci dalam menjembatani hambatan komunikasi lintas bahasa secara otomatis, baik dalam konteks akademik, industri, maupun layanan publik. Perkembangan MT menjadi semakin pesat seiring meningkatnya kebutuhan pertukaran informasi global yang cepat dan akurat [1]. Dalam satu dekade terakhir, pendekatan *Neural Machine Translation* (NMT) dengan kerangka kerja *end-to-end* telah mendominasi riset dan aplikasi MT [2]. Dominasi ini didorong oleh kemampuan NMT dalam memodelkan relasi linguistik yang kompleks secara lebih efektif dibandingkan metode konvensional seperti Statistical Machine Translation (SMT) [3].

Keberhasilan NMT sangat menonjol pada bahasa dengan sumber daya tinggi (*High-Resource Languages* atau HRL), yang didukung oleh ketersediaan korpora paralel berskala besar. Namun, kondisi ini berbanding terbalik pada bahasa dengan

Received: 21-02-2026 | Revised: 27-06-2026 | Accepted: 03-07-2026 | Published: 15-07-2026

sumber daya rendah (*Low-Resource Languages* atau LRL), di mana keterbatasan data paralel menjadi hambatan utama dalam pelatihan model NMT yang andal [4]. Sejumlah studi menunjukkan bahwa ketika jumlah data pelatihan berada pada skala terbatas, misalnya di bawah satu juta pasangan kalimat kinerja NMT cenderung menurun drastis dan dalam beberapa kasus sulit melampaui performa SMT [5]. Oleh karena itu, tantangan sentral dalam pengembangan NMT untuk LRL terletak pada bagaimana memaksimalkan pembelajaran model di tengah kondisi kelangkaan data.

Untuk mengatasi keterbatasan tersebut, penelitian mutakhir banyak mengalihkan perhatian pada pemanfaatan data monolingual yang relatif lebih melimpah. Dua pendekatan yang paling dominan dalam konteks ini adalah *Transfer Learning* (TL) dan *Data Augmentation* (DA). TL berupaya mentransfer pengetahuan dari bahasa atau domain yang kaya sumber daya ke skenario LRL melalui mekanisme *pre-training* dan *fine-tuning* [6].
Sementara itu, DA berfokus pada perluasan data pelatihan melalui sintesis data buatan, seperti *Back-Translation* (BT), sehingga model dapat belajar dari variasi data yang lebih beragam meskipun data paralel asli terbatas [7].

Beberapa studi tinjauan literatur (review dan survei) sebelumnya telah berupaya memetakan perkembangan NMT. Sebagai contoh, Tan et al. [2] memberikan tinjauan NMT secara umum, namun cakupannya terlalu luas dan belum berfokus pada penanganan skenario sumber daya rendah menggunakan arsitektur terkini. Di ranah bahasa sumber daya rendah, Zhang et al. [8] menyajikan survei NMT, namun kajian tersebut lebih berpusat pada perspektif bahasa Tiongkok dan belum menyajikan tinjauan sistematis secara global. Sementara itu, Tafa et al. [9] telah melakukan tinjauan sistematis mengenai kinerja penerjemahan mesin untuk bahasa sumber daya rendah secara umum. Meskipun studi-studi tinjauan tersebut sangat berharga dalam memetakan performa metrik evaluasi, belum ada tinjauan sistematis yang secara khusus dan komprehensif membedah integrasi metodologis antara strategi hibrida (Hybrid Transfer Learning) dengan pendekatan objektif self-supervised (seperti MLM dan DAE) sebagai satu ekosistem solusi yang utuh.

Lebih lanjut, meskipun efektivitas TL dan DA telah banyak dilaporkan, berbagai kajian literatur menunjukkan bahwa sebagian besar penelitian masih cenderung mengeksplorasi kedua pendekatan tersebut secara terpisah. Sebagai contoh, sejumlah tinjauan dan studi primer memfokuskan analisis pada BT sebagai teknik DA utama tanpa mengaitkannya dengan strategi TL yang lebih luas, sementara penelitian lain menitikberatkan pada inisialisasi pre-training berbasis bahasa sumber daya tinggi tanpa mengevaluasi kontribusi augmentasi data secara bersilangan. Di samping itu, integrasi objektif self-supervised—seperti *Masked Language Modeling* (MLM), *Causal Language Modeling* (CLM), dan *Denoising Autoencoder* (DAE) dengan strategi hibrida masih relatif jarang dikaji secara terpadu, meskipun pendekatan ini menunjukkan potensi besar dalam meningkatkan kualitas representasi linguistik pada skenario LRL. Ketiadaan tinjauan sistematis yang mengontraskan dan mensintesis integrasi pendekatan-pendekatan inilah yang menimbulkan kesenjangan pemahaman mengenai tren, efektivitas relatif, serta potensi sinergi antar-metode dalam pengembangan NMT untuk LRL [10].

Berdasarkan latar belakang tersebut, penelitian ini bertujuan menyajikan tinjauan literatur sistematis mengenai penerapan teknik *Transfer Learning* dan *Data Augmentation* dalam meningkatkan kinerja NMT pada LRL. Kajian ini mencakup identifikasi pendekatan mutakhir, evaluasi efektivitas masing-masing teknik, serta analisis terhadap kombinasi TL dan DA, termasuk pemanfaatan objektif *self-supervised*. Secara khusus, tinjauan ini diharapkan dapat (1) menyajikan peta komprehensif pendekatan TL dan DA yang digunakan dalam NMT LRL, (2) merumuskan tren implementasi dan integrasi kedua strategi tersebut, serta (3) memberikan kerangka konseptual yang dapat menjadi landasan bagi penelitian lanjutan dan pengembangan sistem NMT yang lebih efektif pada skenario sumber daya rendah.

Berdasarkan kesenjangan penelitian yang telah diidentifikasi, tinjauan literatur sistematis ini dirancang untuk menjawab tiga pertanyaan penelitian utama. Pertama, penelitian ini menelaah bagaimana teknik *Transfer Learning* (TL) dan *Data Augmentation* (DA) secara individual berkontribusi dalam mengatasi keterbatasan korpus paralel pada skenario NMT bahasa sumber daya rendah (RQ1). Kedua, penelitian ini mengkaji sejauh mana integrasi TL dan DA, termasuk pendekatan *Hybrid Transfer Learning* serta pemanfaatan objektif *self-supervised* seperti *Masked Language Modeling*, *Causal Language Modeling*, dan *Denoising Autoencoder*, menunjukkan kecenderungan peningkatan kinerja yang lebih stabil dan efektif dibandingkan penerapan tunggal masing-masing teknik (RQ2). Ketiga, penelitian ini mengidentifikasi tantangan metodologis utama, keterbatasan yang masih melekat, serta peluang dan arah penelitian masa depan dalam penerapan gabungan TL dan DA untuk pengembangan NMT pada lingkungan sumber daya rendah (RQ3).

Meskipun beberapa survei telah membahas Transfer Learning maupun Data Augmentation secara terpisah, perkembangan terbaru menunjukkan bahwa sebagian besar model NMT modern untuk bahasa sumber daya rendah justru mengombinasikan kedua strategi tersebut bersama teknik self-supervised learning. Oleh karena itu, diperlukan sintesis yang tidak hanya membandingkan masing-masing pendekatan, tetapi

juga menjelaskan bagaimana ketiganya saling melengkapi dalam membangun sistem NMT yang lebih efektif. Kontribusi utama penelitian ini adalah**:** (1) menyusun sintesis sistematis mengenai perkembangan Transfer Learning dan Data Augmentation pada NMT bahasa sumber daya rendah; (2) mengidentifikasi tren integrasi kedua pendekatan dengan teknik self-supervised learning; serta (3) merumuskan arah penelitian masa depan berdasarkan kesenjangan yang ditemukan pada studi-studi terkini.

2. Metodologi Penelitian

Tinjauan literatur sistematis ini difokuskan pada karya ilmiah yang membahas teknik-teknik untuk mengatasi masalah sumber daya rendah dalam Penerjemahan Mesin Neural. Ruang lingkup tematik yang diidentifikasi dari sumber-sumber yang tersedia meliputi: Transfer Learning (TL), termasuk teknik-teknik seperti *Hybrid Transfer Learning* (HTL); Pemanfaatan Data Monolingual melalui *Back-Translation* (BT), *Self-Training*, dan objektif *self-supervised* seperti *Masked Language Modeling* (MLM), *Causal Language Modeling* (CLM), dan *Denoising Autoencoder* (DAE); dan aplikasi NMT pada pasangan bahasa LRL (misalnya, Azeri-Tiongkok, Uzbek-Tiongkok, Khmer-Vietnam, Tiongkok-Urdu, Tiongkok-Vietnam).

2.1. Desain Penelitian

Penelitian ini dirancang sebagai tinjauan literatur sistematis (*Systematic Literature Review*) yang bertujuan untuk mengidentifikasi, mengevaluasi, dan menyintesis karya ilmiah terkait strategi peningkatan kualitas *Neural Machine Translation* (NMT) pada bahasa sumber daya rendah (*Low-Resource Languages*/LRL). Proses tinjauan ini mengadopsi pedoman standar *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) 2020 untuk memastikan transparansi dan replikabilitas studi. Fokus utama tinjauan mencakup teknik *Transfer Learning* (TL), Augmentasi Data (DA), serta pemanfaatan objektif *self-supervised*.

2.2. Sumber Data dan Database

Pencarian literatur dilakukan menggunakan Scopus sebagai basis data ilmiah utama. Pemilihan Scopus didasarkan pada cakupannya yang luas terhadap literatur akademik berkualitas tinggi yang relevan dengan topik ilmu komputer dan linguistik komputasional. Proses pencarian awal dilakukan pada tanggal 20 Oktober 2025, dengan pembaruan pencarian dilakukan pada 11 Desember 2025 untuk memastikan inklusi studi terbaru.

2.3. Strategi Pencarian (Search Strategy)

Strategi pencarian literatur dirancang menggunakan pendekatan multi-string untuk memastikan cakupan yang komprehensif terhadap dua domain utama penelitian, yaitu teknik augmentasi data dan metode transfer pembelajaran. Pencarian dilakukan pada basis data Scopus dengan dua string kueri yang saling melengkapi:

1. Kueri A (Fokus Augmentasi Data): Dirancang untuk menjaring variasi teknik augmentasi, khususnya *Back-Translation*. TITLE-ABS-KEY (back translation) AND PUBYEAR > 2020 AND PUBYEAR < 2027 AND (LIMIT-TO (OA , "all")) AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (LANGUAGE , "english")).
2. Kueri B (Fokus Transfer Learning & NMT): Dirancang untuk menangkap metode inisialisasi model dan konteks sumber daya rendah secara spesifik. TITLE-ABS-KEY ("neural machine translation" AND "low resource" AND ("transfer learning" OR "pre-training")) AND PUBYEAR > 2020 AND PUBYEAR < 2027 AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (LANGUAGE , "english")) AND (LIMIT-TO (OA , "all")).

Kombinasi kedua pencarian ini menghasilkan total 1.688 artikel (1.661 dari Kueri A dan 27 dari Kueri B) sebelum proses penyaringan.

2.4. Kriteria Inklusi dan Eksklusi

Untuk menjamin relevansi studi yang dipilih, kriteria inklusi dan eksklusi diterapkan secara ketat dalam dua tahap penyaringan:

Kriteria Inklusi:

- Artikel jurnal ("ar") yang dipublikasikan dalam rentang tahun 2021 hingga 2026 (sesuai batasan tahun kueri >2020).
- Artikel tersedia dalam akses terbuka (*Open Access*) dan ditulis dalam Bahasa Inggris.
- Artikel secara eksplisit membahas konteks bahasa sumber daya rendah (*low-resource languages*).
- Artikel membahas implementasi *Transfer Learning*, *Data Augmentation* (termasuk *Back-Translation*), atau objektif *self-supervised*.

Kriteria Eksklusi:

- Artikel yang hanya menyebutkan istilah "back translation" secara sekilas tanpa menjadikannya fokus penelitian atau metode utama (Tahap 1).

- Artikel yang tidak secara eksplisit membahas atau menguji metode pada konteks bahasa sumber daya rendah (Tahap 2).

2.5. Proses Screening dan Seleksi Studi

Proses seleksi studi dilakukan melalui penyaringan bertingkat untuk menjamin relevansi dan validitas metodologis. Mengingat tingginya jumlah artikel awal, strategi penyaringan difokuskan pada eliminasi ambiguitas istilah dan pengetatan konteks komputasi.

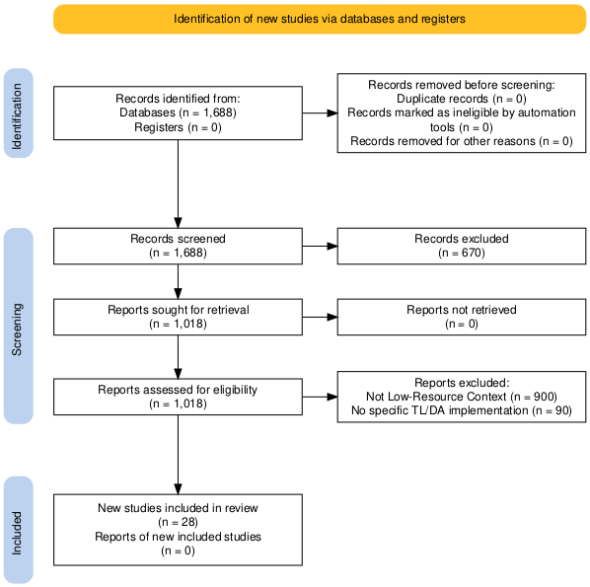
1. Identifikasi dan Validasi Awal (n = 1.688) Tahap identifikasi awal mengungkapkan adanya tantangan polisemi pada kata kunci "back translation". Analisis pendahuluan menunjukkan bahwa sebagian besar hasil dari Kueri A merujuk pada prosedur validasi instrumen psikometri (kuesioner) di bidang medis dan ilmu sosial, bukan teknik komputasi. Sebaliknya, hasil dari Kueri B menunjukkan relevansi yang sangat tinggi (presisi) terhadap topik NMT.

2. Penyaringan Tahap 1: Eksklusi Domain Non-Komputasi (from 1.688 to 1.018). Penyaringan manual dilakukan berdasarkan judul dan abstrak untuk memisahkan studi komputasi dari studi klinis/sosial.

- Kriteria Eksklusi: Artikel yang membahas adaptasi lintas budaya (*cross-cultural adaptation*), uji klinis, atau analisis sentimen tanpa fokus pada mekanisme penerjemahan dieksklusi.
- Langkah ini secara signifikan mereduksi *false positives* dari pencarian awal, menyisakan 1.018 artikel yang relevan secara teknis (termasuk seluruh 27 artikel dari Kueri B).

3. Penyaringan Tahap 2: Kelayakan Konteks (*Eligibility*) (n = 1.018 n = 28) Artikel yang tersisa ditinjau secara mendalam (*full-text review*) dengan kriteria inklusi ketat: fokus pada *Low-Resource Languages* (LRL) dan penerapan metode NMT.

- Sebagian besar artikel dari Kueri A dieksklusi pada tahap ini karena berfokus pada bahasa sumber daya tinggi (*High-Resource*) atau hanya menggunakan *back-translation* sebagai standar *baseline* tanpa analisis mendalam.
- Sebaliknya, mayoritas artikel dari Kueri B dipertahankan karena secara eksplisit membahas strategi penanganan kelangkaan data.
- Setelah menghapus duplikasi antar-kueri dan mengeluarkan studi yang tidak memenuhi standar (misalnya, topik klasifikasi teks atau *sentiment analysis*), diperoleh 28 studi final yang dimasukkan ke dalam tinjauan sistematis ini.



Gambar 1. Diagram alir PRISMA 2020 yang merangkum proses identifikasi, penyaringan, dan seleksi studi [11]

2.6. Proses Ekstraksi Data

Data dari 28 studi final diekstraksi secara manual menggunakan formulir ekstraksi data terstruktur untuk memastikan konsistensi dan akurasi pencatatan. Proses ekstraksi difokuskan pada variabel-variabel kunci yang relevan untuk menjawab pertanyaan penelitian, yang dikelompokkan sebagai berikut:

1. Informasi Bibliografi: Mencakup penulis, tahun publikasi, judul, dan jenis publikasi (jurnal/konferensi) untuk memetakan distribusi temporal dan demografi penelitian.
2. Karakteristik Linguistik & Dataset: Mencatat pasangan bahasa spesifik yang diteliti (misalnya, *Kazakh-Inggris*, *Dongba-Tiongkok*, atau *Batak Toba-Indonesia*) serta skala data yang digunakan (jumlah kalimat paralel dan monolingual). Informasi ini krusial untuk memvalidasi status "sumber daya rendah" dari studi kasus yang dipilih.
3. Klasifikasi Metodologi: Setiap studi dikategorikan berdasarkan pendekatan teknis utama:
    - Transfer Learning (TL): Termasuk *pre-training* model bahasa (misal: mBART), *Hybrid Transfer Learning* (HTL), dan strategi inisialisasi parameter.
    - Augmentasi Data (DA): Termasuk teknik *Back-Translation* (BT), *Pivot-Based Augmentation*, dan pembangkitan data sintetis.
4. Metrik Evaluasi: Mencatat metrik standar yang digunakan untuk mengukur kualitas terjemahan, terutama skor BLEU (*Bilingual Evaluation Understudy*), serta metrik pendukung seperti chrF dan TER jika tersedia.

5. Hasil Utama: Mengekstraksi temuan kunci berupa perbandingan kinerja kuantitatif (delta peningkatan skor terhadap *baseline*) serta analisis kualitatif mengenai efektivitas metode yang diusulkan.

Seluruh data yang diekstraksi kemudian disintesis secara naratif untuk menganalisis tren keberhasilan masing-masing metode dalam mengatasi masalah kelangkaan data.

Untuk memastikan validitas temuan, penilaian kualitas (Quality Assessment) dilakukan pada 28 studi final. Kriteria kualitas dievaluasi berdasarkan tiga metrik utama: (1) kejelasan arsitektur eksperimen (termasuk deskripsi baseline model yang digunakan), (2) transparansi penggunaan dataset (skala korpus paralel maupun monolingual), dan (3) pelaporan metrik evaluasi standar yang dapat diandalkan, seperti BLEU atau chrF. Seluruh studi yang disertakan dalam tinjauan ini telah memenuhi ambang batas kualitas eksperimen komputasi yang memadai.

2.7. Proses Sintesis Data (Synthesis Method)

Data yang diekstraksi kemudian disintesis secara kualitatif-naratif untuk menjawab pertanyaan penelitian yang telah dirumuskan. Sintesis ini mengelompokkan temuan berdasarkan pendekatan metodologis utama—yaitu *Transfer Learning*, Augmentasi Data, dan pendekatan Hibrida—untuk menganalisis efektivitas, tren implementasi, dan sinergi antar-metode dalam meningkatkan kualitas terjemahan pada skenario sumber daya rendah [12], [13], [14], [15], [16] [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36].

3. Hasil dan Pembahasan

3.1. Sintesis Strategi Penanganan Kelangkaan Data

Berdasarkan tinjauan terhadap 28 studi terpilih, strategi peningkatan kinerja NMT pada bahasa sumber daya rendah (*Low-Resource Languages*/LRL) tidak lagi berdiri sebagai teknik yang terisolasi, melainkan membentuk sebuah ekosistem solusi yang saling melengkapi. Analisis kami mengklasifikasikan pendekatan-pendekatan tersebut ke dalam tiga pilar utama yang bekerja secara sinergis: (1) Inisialisasi Pengetahuan melalui *Transfer Learning* (TL), (2) Ekspansi Datamelalui *Data Augmentation* (DA), dan (3) Integrasi Hibrida (*Hybrid Transfer Learning*/HTL) yang menjembatani kesenjangan leksikal antara bahasa sumber dan target.

Secara konseptual, hubungan antar-strategi ini dapat diringkas dalam Tabel 1, yang menunjukkan pergeseran tren dari penggunaan metode tunggal menuju pendekatan gabungan yang lebih *robust*.

3.2. Transfer Learning: Dari Inisialisasi Parameter hingga Adaptasi Leksikon

Mayoritas studi menegaskan bahwa *Transfer Learning* (TL) merupakan langkah inisialisasi wajib untuk skenario LRL. Teknik dominan melibatkan pelatihan awal (*pre-training*) pada korpus bahasa sumber daya tinggi (HRL) sebelum dilakukan *fine-tuning* ke bahasa target. Studi terbaru menunjukkan bahwa pemanfaatan Model Bahasa Besar (LLMs) multikual seperti mBART dan T5 memberikan lonjakan kinerja yang signifikan dibandingkan inisialisasi acak. Secara spesifik, pendekatan *Mutual Transfer* dan *Inverted Transfer* mulai diadopsi, di mana bahasa dengan kerabat dekat (misalnya Turki-Azeri atau Melayu-Indonesia) saling berbagi parameter untuk memperkaya representasi fitur. Meskipun TL terbukti efektif, analisis mendalam mengungkap adanya tantangan kesenjangan leksikal (*lexical gap*). Model induk (HRL) seringkali gagal mengenali entitas spesifik pada bahasa anak (LRL) karena perbedaan morfologi yang tajam. Gap Teridentifikasi: TL konvensional hanya mentransfer bobot *encoder-decoder*, namun sering mengabaikan adaptasi *embedding* kosakata. Implikasi: Inilah mengapa metode seperti yang diusulkan Maimaiti et al. (2022) menjadi krusial, di mana *pre-trained lexicon embedding* disuntikkan secara eksplisit. Hal ini mengindikasikan bahwa masa depan TL pada LRL bukan lagi sekadar "copy-paste" parameter, melainkan adaptasi representasi semantik yang lebih *fine-grained.*

3.3. Data Augmentation: Menyeimbangkan Kuantitas dan Kualitas Data Sintetis

Back-Translation (BT) tetap menjadi teknik augmentasi paling populer (*de facto standard*). Temuan dari Kimera et al. (2025) pada pasangan Inggris-Luganda dan Quoc et al. (2023) pada Khmer-Vietnam menunjukkan bahwa BT secara konsisten meningkatkan skor BLEU dengan memanfaatkan data monolingual yang melimpah. Selain itu, strategi berbasis *Pivot* (menggunakan Bahasa Inggris sebagai perantara) terbukti efektif untuk pasangan bahasa yang sangat jarang (*zero-resource*), seperti yang ditemukan pada kasus bahasa daerah di Tiongkok (Dongba). Namun, pola menarik ditemukan terkait batas saturasi data. Vu & Bui (2023) memperingatkan adanya *diminishing returns* atau bahkan penurunan kinerja ketika rasio data sintetis terlalu mendominasi data asli (misalnya rasio > 1:1 atau 2:1).

Tabel 1. Ringkasan Sintesis Pendekatan NMT pada LRL

Kategori Strategi	Peran Utama	Temuan Kunci (*Result*)	Analisis Kritis (*Analysis*)
Transfer Learning (TL)	*Foundation* (Pondasi Awal)	Penggunaan model *pre-trained* (mBART, MarianMT) meningkatkan konvergensi model.	Efektivitas sangat bergantung pada kedekatan linguistik; rentan terhadap *negative transfer* jika domain atau bahasa terlalu berbeda.

Data Augmentation (DA)	Expansion (*Volume Data*)	Back-Translation (BT) dan Pivotmen ambah variasi data pelatihan secara signifikan.	Kuantitas tidak selalu berbanding lurus dengan kualitas; saturasi data sintetis dapat memicu "halusinasi" model tanpa filter yang tepat.
Hybrid & Self-Supervised	*Bridge* (Jembatan)	Menggabungkan TL + DA dengan objektif self-supervised (MLM/CLM) menghasilkan skor BLEU tertinggi.	Solusi paling efektif untuk *extremely low-resource*, karena mengatasi kelemahan TL (*mismatch* kosakata) dan DA (*noise*).

Penambahan data sintetik tanpa kontrol kualitas menyebabkan model "melupakan" distribusi data asli yang benar (*catastrophic forgetting*) dan meningkatkan *noise*.

Munculnya tren *Tagged Back-Translation* (memberi penanda khusus <BT> pada data sintetis) menjadi solusi kritis. Tag ini memungkinkan model untuk membedakan sumber data, "mempercayai" data asli lebih tinggi daripada data sintetis, namun tetap belajar dari variasi linguistik data sintetis. Ini menunjukkan pergeseran fokus dari "memperbanyak data" menjadi "mengelola kualitas data".

3.4. Hybrid Transfer Learning (HTL) dan Integrasi Self-Supervised: Menuju Ekosistem Terpadu

Temuan paling signifikan dari tinjauan ini adalah superioritas pendekatan hibrida. Studi seperti Tafa et al. (2025) dan Yazar & Kilic (2025) melaporkan bahwa kombinasi TL dan DA menghasilkan peningkatan kinerja yang lebih tinggi dibandingkan penjumlahan peningkatan masing-masing metode secara terpisah. Lebih lanjut, integrasi objektif *self-supervised* seperti *Masked Language Modeling* (MLM) dan *Denoising Autoencoder* (DAE) ke dalam alur pelatihan hibrida mampu mengekstrak struktur sintaksis dari data monolingual sebelum proses penerjemahan dimulai. Keunggulan HTL terletak pada kemampuannya menutupi kelemahan masing-masing komponen pembentuknya.

TL menyediakan inisialisasi parameter yang baik (mengatasi masalah "cold start" pada model neural), sementara DA menyediakan volume data yang cukup untuk mencegah *overfitting* saat *fine-tuning*. HTL dengan objektif *self-supervised* (MLM/DAE) menciptakan model yang tidak hanya "menerjemahkan" tetapi juga "memahami" struktur bahasa. Dalam konteks *extremely low-resource* (<10.000 kalimat), mengandalkan satu teknik saja tidak lagi memadai. Masa depan pengembangan NMT untuk LRL terletak pada arsitektur integratif yang mampu melakukan *transfer learning* fitur leksikon sekaligus melakukan augmentasi data secara cerdas dan terukur.

Penelitian ini memiliki beberapa keterbatasan terkait strategi pencarian literatur. Pertama, pencarian dibatasi pada basis data Scopus dan publikasi Open Access untuk memastikan aksesibilitas dan replikabilitas studi. Hal ini berpotensi menimbulkan selection bias, di mana studi-studi relevan yang diterbitkan di luar indeks Scopus (misalnya ACL Anthology) atau jurnal berbayar (closed-access) mungkin tidak tercakup. Untuk penelitian mendatang, disarankan untuk memperluas pencarian melintasi berbagai repositori khusus ilmu komputer dan linguistik komputasional.

4. Kesimpulan

Masalah Penelitian ini telah menyajikan tinjauan literatur sistematis terhadap 28 studi terkini mengenai strategi penanganan kelangkaan data dalam *Neural Machine Translation* (NMT) untuk bahasa sumber daya rendah. Temuan utama menegaskan bahwa paradigma pengembangan NMT telah bergeser secara fundamental dari penerapan teknik tunggal menuju ekosistem solusi yang terintegrasi penuh. Analisis mendalam menunjukkan bahwa *Transfer Learning* (TL) dan *Data Augmentation* (DA) memiliki peran komplementer yang vital dalam arsitektur model modern. TL berfungsi krusial sebagai inisialisasi parameter untuk mengatasi hambatan awal pelatihan (*cold-start*), sementara DA berperan signifikan dalam mengekspansi variasi data guna mencegah *overfitting*.

Secara spesifik, pendekatan *Hybrid Transfer Learning* (HTL) yang mengintegrasikan inisialisasi model pra-latih dengan strategi augmentasi terbukti menghasilkan kinerja yang paling superior dan stabil dibandingkan metode lainnya. Sinergi ini efektif menutupi kelemahan parsial masing-masing teknik, di mana augmentasi mengisi celah leksikal yang sering terlewat oleh TL, sedangkan TL menjaga integritas struktur sintaksis dari gangguan data sintetis. Namun, analisis juga menggarisbawahi bahwa penggunaan data sintetis tanpa kontrol kualitas rentan terhadap saturasi yang dapat memicu penurunan akurasi model. Oleh karena itu, integrasi objektif *self-supervised* seperti *Masked Language Modeling* (MLM) menjadi mekanisme esensial untuk memfilter *noise* dan memperkuat pemahaman semantik model sebelum proses penerjemahan.

Merespons tantangan yang tersisa, arah penelitian masa depan perlu difokuskan pada pengembangan teknik augmentasi cerdas yang memprioritaskan kualitas data dibandingkan sekadar kuantitas semata. Peneliti juga didorong untuk mengeksplorasi metode adaptasi leksikon yang lebih efisien secara komputasi agar kosakata bahasa lokal dapat disuntikkan ke model besar tanpa biaya pelatihan ulang yang mahal. Validasi terhadap model hibrida ini harus diperluas ke domain spesifik yang lebih teknis untuk menjamin kebermanfaatan praktis sistem penerjemahan di

berbagai bidang keilmuan. Akhirnya, tinjauan ini menyimpulkan bahwa orkestrasi yang tepat antara transfer pengetahuan global dan pemanfaatan data lokal merupakan kunci utama dalam menjembatani kesenjangan komunikasi digital dunia.

Reference

[1] D. Ataman, A. Birch, N. Habash, M. Federico, P. Koehn, and K. Cho, "Machine Translation in the Era of Large Language Models:A Survey of Historical and Emerging Problems," *Information 2025, Vol. 16, Page 723*, vol. 16, no. 9, p. 723, Aug. 2025, doi: 10.3390/INFO16090723.

[2] Z. Tan et al., "Neural machine translation: A review of methods, resources, and tools," AI Open, vol. 1, pp. 5–21, Jan. 2020, doi: 10.1016/J.AIOPEN.2020.11.001.

[3] A. Ramesh, V. B. Parthasarathy, R. Haque, and A. Way, "Comparing Statistical and Neural Machine Translation Performance on Hindi-To-Tamil and English-To-Tamil," Digital 2021, Vol. 1, Pages 86-102, vol. 1, no. 2, pp. 86–102, Apr. 2021, doi: 10.3390/DIGITAL1020007.

[4] A. Solyman et al., "Optimizing the impact of data augmentation for low-resource grammatical error correction," Journal of King Saud University - Computer and Information Sciences, vol. 35, no. 6, Jun. 2023, doi: 10.1016/j.jksuci.2023.101572.

[5] S. Sen, M. Hasanuzzaman, A. Ekbal, P. Bhattacharyya, and A. Way, "Neural machine translation of low-resource languages using SMT phrase pair injection," Nat. Lang. Eng., vol. 27, no. 3, pp. 271–292, May 2021, doi: 10.1017/S1351324920000303.

[6] J. Dong, "Transfer Learning-Based Neural Machine Translation for Low-Resource Languages," ACM Transactions on Asian and Low-Resource Language Information Processing, Sep. 2023, doi: 10.1145/3618111.

[7] R. Kimera, D. N. Heo, D. N. Rim, and H. Choi, "Data Augmentation With Back translation for Low Resource languages: A case of English and Luganda," in NLPIR 2024 - 2024 8th International Conference on Natural Language Processing and Information Retrieval, Association for Computing Machinery, Inc, Apr. 2025, pp. 142–148. doi: 10.1145/3711542.3711594.

[8] J. Zhang et al., "Neural Machine Translation for Low-Resource Languages from a Chinese-centric Perspective: A Survey," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 23, no. 6, Jun. 2024, doi: 10.1145/3665244.

[9] T. O. Tafa et al., "Machine Translation Performance for Low-Resource Languages: A Systematic Literature Review," IEEE Access, vol. 13, pp. 72486–72505, 2025, doi: 10.1109/ACCESS.2025.3562918.

[10] W.-H. Her and U. Kruschwitz, "Investigating Neural Machine Translation for Low-Resource Languages: Using Bavarian as a Case Study," Apr. 2024, Accessed: Nov. 25, 2025. [Online]. Available: http://arxiv.org/abs/2404.08259

[11] N. R. Haddaway, M. J. Page, C. C. Pritchard, and L. A. McGuinness, "PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis," Campbell Systematic Reviews, vol. 18, no. 2, p. e1230, Jun. 2022, doi: 10.1002/CL2.1230.

[12] A. Javed et al., "Transformer-Based Re-Ranking Model for Enhancing Contextual and Syntactic Translation in Low-Resource Neural Machine Translation," Electronics (Switzerland), vol. 14, no. 2, Jan. 2025, doi: 10.3390/electronics14020243.

[13] J. Pang et al., "Rethinking the Exploitation of Monolingual Data for Low-Resource Neural Machine Translation under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license," Computational Linguistics, vol. 50, no. 1, 2024, doi: 10.1162/coli.

[14] T. N. Quoc, H. Le Thanh, and H. P. Van, "Khmer-Vietnamese Neural Machine Translation Improvement Using Data Augmentation Strategies," Informatica (Slovenia), vol. 47, no. 3, pp. 349–360, Sep. 2023, doi: 10.31449/inf.v47i3.4761.

[15] H. Vu and N. D. Bui, "On the scalability of data augmentation techniques for low-resource machine translation between Chinese and Vietnamese," Journal of Information and Telecommunication, vol. 7, no. 2, pp. 241–253, 2023, doi: 10.1080/24751839.2023.2186625.

[16] M. Maimaiti, Y. Liu, H. Luan, and M. Sun, "Enriching the Transfer Learning with Pre-Trained Lexicon Embedding for Low-Resource Neural Machine Translation," 2022.

[17] C. Samuel and I. T. Ali, "Batak Toba language-Indonesian machine translation with transfer learning using no language left behind," International Journal of Advances in Applied Sciences, vol. 13, no. 4, pp. 830–839, Dec. 2024, doi: 10.11591/ijaas.v13.i4.pp830-839.

[18] L. Li, W. Hu, and M. Luo, "PNMT: Zero-Resource Machine Translation with Pivot-Based Feature Converter," Computers, Materials and Continua, vol. 84, no. 3, pp. 5915–5935, 2025, doi: 10.32604/cmc.2025.064349.

[19] X. Wu and R. Deng, "Research on the application of cross-language transfer learning model in English translation for low-resource scenarios," IEEE Access, 2025, doi: 10.1109/ACCESS.2025.3628643.

[20] J. Pang et al., "Salute the Classic: Revisiting Challenges of Machine Translation in the Age of Large Language Models", doi: 10.1162/tacl.

[21] B. K. Yazar and E. Kilic, "Improving Low-Resource Kazakh-English and Turkish-English Neural Machine Translation Using Transfer Learning and Part of Speech Tags," IEEE Access, vol. 13, pp. 32341–32356, 2025, doi: 10.1109/ACCESS.2025.3542491.

[22] K. Ismail, S. Abdou, M. Farouk, and A. Salem, "Transformers to the rescue: alleviating data scarcity in arabic grammatical error correction with pre-trained models," Neural Comput. Appl., vol. 37, no. 18, pp. 13011–13038, Jun. 2025, doi: 10.1007/s00521-025-11145-1.

[23] V. Lara-Ortiz, R. Q. Fuentes-Aguilar, and I. Chairez, "Spanish to Mexican Sign Language glosses corpus for natural language processing tasks," Scientific Data , vol. 12, no. 1, Dec. 2025, doi: 10.1038/s41597-025-04871-7.

[24] X. Ma, M. Lan, W. Hu, and Y. Lu, "Dongba Machine Translation with Transfer Learning: Leveraging Pre-trained Ancient Chinese Models," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 24, no. 5, Apr. 2025, doi: 10.1145/3721980.

[25] M. R. Costa-jussà et al., "Scaling neural machine translation to 200 languages," Nature, vol. 630, no. 8018, pp. 841–846, Jun. 2024, doi: 10.1038/s41586-024-07335-x.

[26] W. Zhang, X. Li, Y. Yang, R. Dong, and V. Basile, "Pre-Training on Mixed Data for Low-Resource Neural Machine Translation," 2021, doi: 10.3390/info1203.

[27] M. Tars, A. Tättar, and M. Fishel, "Cross-lingual Transfer from Large Multilingual Translation Models to Unseen Under-resourced Languages," Baltic Journal of Modern Computing, vol. 10, no. 3, pp. 435–446, 2022, doi: 10.22364/bjmc.2022.10.3.16.

[28] J. Yan, T. Lin, and S. Zhao, "Migration Learning and Multi-View Training for Low-Resource Machine Translation Migration Learning and Multi-View Training," 2024. [Online]. Available: www.ijacsa.thesai.org

[29] Y. Wen, J. Guo, Z. Yu, and Z. Yu, "Chinese–Vietnamese Pseudo-Parallel Sentences Extraction Based on Image Information Fusion," Information (Switzerland), vol. 14, no. 5, May 2023, doi: 10.3390/info14050298.

[30] A. Slim, A. Melouah, U. Faghihi, and K. Sahib, "Improving Neural Machine Translation for Low Resource Algerian Dialect by Transductive Transfer Learning Strategy," Arab. J. Sci. Eng., vol. 47, no. 8, pp. 10411–10418, Aug. 2022, doi: 10.1007/s13369-022-06588-w.

[31] T. V. Ngo, P. T. Nguyen, V. V. Nguyen, T. Le Ha, and L. M. Nguyen, "An Efficient Method for Generating Synthetic Data for Low-Resource Machine Translation: An empirical study of Chinese, Japanese to Vietnamese Neural Machine

Translation," Applied Artificial Intelligence, vol. 36, no. 1, 2022, doi: 10.1080/08839514.2022.2101755.
[32] Z. Xu, S. Zhan, W. Yang, and Q. Xie, "Based on Gated Dynamic Encoding Optimization, the LGE-Transformer Method for Low-Resource Neural Machine Translation," IEEE Access, vol. 12, pp. 162861–162869, 2024, doi: 10.1109/ACCESS.2024.3488186.
[33] M. Sun, H. Wang, M. Pasquine, and I. A. Hameed, "Machine translation in low-resource languages by an adversarial neural network," Applied Sciences (Switzerland), vol. 11, no. 22, Nov. 2021, doi: 10.3390/app112210860.
[34] Y. Sang, Y. Chen, and J. Zhang, "Neural Machine Translation Research on Syntactic Information Fusion Based on the Field of Electrical Engineering," Applied Sciences (Switzerland), vol. 13, no. 23, Dec. 2023, doi: 10.3390/app132312905.
[35] Z. Mao, C. Chu, and S. Kurohashi, "Linguistically Driven Multi-Task Pre-Training for Low-Resource Neural Machine Translation," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 21, no. 4, Jul. 2022, doi: 10.1145/3491065.
[36] V. Karyukin, D. Rakhimova, A. Karibayeva, A. Turganbayeva, and A. Turarbek, "The neural machine translation models for the low-resource Kazakh–English language pair," PeerJ Comput. Sci., vol. 9, 2023, doi: 10.7717/peerj-cs.1224.