Enhancing Translation From Hebrew to Low-Resource Languages Through an English Intermediary Approach

Atalia Solash
Ben-Gurion University
Beer-Sheva, Israel
solash@post.bgu.ac.il

Shir Mashiah
Ben-Gurion University
Beer-Sheva, Israel
shirmash@post.bgu.ac.il

Amit Fridman
Ben-Gurion University
Beer-Sheva, Israel
amitfrid@post.bgu.ac.il

Erica Rusonik
Ben-Gurion University
Beer-Sheva, Israel
ericaru@post.bgu.ac.il

ABSTRACT

Translating between Hebrew and low-resource languages presents significant challenges due to limited parallel corpora and distinct linguistic differences. This study explores the use of pivot-based machine translation, with English as an intermediary, to improve translation quality for Hebrew-Finnish and Hebrew-Ukrainian language pairs. To evaluate this approach, we test four state-of-the-art models—NLLB-200, Helsinki-NLP, Google Translate, and Llama 3.1—on the TED 2020 dataset, utilizing 20,000 sentence pairs for each language pair. Translation quality is assessed using BLEU, METEOR, and COMET metrics. Our results show a mixed impact: while pivot-based translation significantly improves BLEU scores for some models, it reduces them for others. METEOR and COMET scores exhibit minor improvements or remain unchanged. Statistical analysis confirms that these differences are significant ($p < 0.05$). However, the variation in BLEU, METEOR, and COMET scores across models highlights that no single approach consistently outperforms the others. These findings emphasize the importance of selecting the right translation model and evaluation metric based on the specific translation task. This study demonstrates the potential of pivot-based translation for low-resource languages while underscoring the challenges of optimizing translation quality across different models.

KEYWORDS

Machine Translation, Pivot Language, Low-Resource Languages

ACM Reference Format:
Atalia Solash, Shir Mashiah, Amit Fridman, and Erica Rusonik. 2025. Enhancing Translation From Hebrew to Low-Resource Languages Through an English Intermediary Approach. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 10 pages.

1 INTRODUCTION

Despite significant advances in machine translation (MT), translating between low-resource languages remains a persistent challenge. The scarcity of parallel corpora and substantial linguistic differences often limit the effectiveness of direct translation models. One promising approach to mitigate these limitations is pivot-based translation, where a high-resource intermediary language, such as English, serves as a bridge between the source and target languages.

This study explores the effectiveness of pivot-based translation for Hebrew-Finnish (H-F) and Hebrew-Ukrainian (H-U) language pairs, which, to the best of our knowledge, have not been previously studied in the context of pivot-based machine translation. While studies on other language pairs have demonstrated notable improvements using pivot-based methods—such as Kenji et al., who reported BLEU score increases of up to 27 points, and Ahmadnia et al., who found enhanced translation accuracy for Persian-Spanish—it remains unclear whether these benefits extend to Hebrew-low-resource language pairs.

To address this gap, we investigate whether pivot-based translation improves translation quality for Hebrew-Finnish and Hebrew-Ukrainian compared to direct translation and how different MT models perform when employing a pivot-based approach versus direct translation. We hypothesize that pivot-based translation will yield higher fluency and semantic adequacy due to the availability of higher-quality Hebrew-English and English-target language corpora.

To evaluate this, we test four state-of-the-art models—NLLB-200, Helsinki-NLP, Google Translate, and Llama 3.1—on the TED 2020 dataset, using BLEU, METEOR, and COMET metrics for a comprehensive assessment. Our findings reveal mixed results: while some models achieve significant BLEU score improvements, others exhibit declines. METEOR and COMET scores remain largely unchanged or show minor gains, indicating that while pivot-based translation helps certain aspects of translation quality, its effectiveness is model-dependent. Statistical analysis confirms these differences as significant ($p < 0.05$), underscoring the complexity of low-resource translation and the necessity for model-specific evaluations.

This paper provides the first detailed assessment of pivot-based translation for Hebrew-low-resource language pairs. It offers a comparative analysis of direct and pivot-based approaches and identifies both the opportunities and limitations of intermediary strategies. The remainder of this paper is structured as follows: Section 2 reviews related work in pivot-based translation and low-resource MT. Section 3 provides background on MT methods and evaluation

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), A. El Fallah Seghrouchni, Y. Vorobeychik, S. Das, A. Nowe (eds.), May 19 – 23, 2025, Detroit, Michigan, USA.

metrics. Section 4 details our methodology and experimental setup. Section 5 presents our results and analysis, and Section 6 concludes with insights and suggestions for future research.

2 RELATED WORK

A lot of research has been conducted on triangulated pivoting, a method in which translation between two languages is achieved by first translating the source language into a high-resource language, and then translating from the pivot language to the target language. Kenji et al. [8] analyzed the effect of using English as the pivot language on translation using a multilingual pre-trained model for translation. They conducted experiments translating Japanese into various low-resource Asian languages using the mBART-50 model [18], extending its word embeddings to include 109 languages from the CC-100 corpus using the method proposed by Wang et al. [24]. Kenji et al. showed that using a pivot language improved, on average, zero-shot translations (translation without training on that language pair) from a BLEU score of 0.5 to 27 when translating Japanese to other languages, and from 0.1 to 17.3 when translating in the opposite direction.

Zhang et al. [26] explored the use of English as a pivot language to improve the instruction-following capabilities of large language models, particularly LLaMA-2. They introduced the pivot language guided generation method, which trains models to process instructions in English before generating responses in the target language. Specifically, the model is trained to first generate an intermediate instruction and response in the pivot language, which then guides the generation of the final response in the target language. This approach, evaluated using the X-AlpacaEval benchmark across four distinct target languages—Chinese, Korean, Italian, and Spanish—demonstrated a notable average improvement of 32% in instruction-following ability for LLaMA-2 compared to the commonly used direct monolingual training method.

Ahmadnia et al. [1] explored the use of English as a pivot language for Persian–Spanish Statistical Machine Translation (SMT), evaluating three translation techniques: direct translation, sentence-level pivoting, and phrase-level pivoting. In direct translation, the process involves translating each word individually. Alternatively, sentence-level pivoting translates entire sentences as cohesive units, while phrase-level pivoting focuses on translating chunks or phrases, rather than individual words, to maintain contextual coherence. In addition they tested two different systems: System (1), which consisted of 10,000 sentences, and System (2), which was expanded to nearly 60,000 sentences by incorporating additional data. The smaller dataset in System (1) simulated a scenario with limited source–target corpora, typical of low-resource languages, while the larger dataset in System (2) simulated high-resource languages, enabling more robust training. This comparison highlighted the significant impact of dataset size on translation performance. Their evaluations, using BLEU scores, revealed that pivot language techniques outperformed direct translation. For System(2), direct translation achieved a BLEU score of 19.39, compared to 21.55 with phrase-level pivoting and 20.78 with sentence-level pivoting. System(1) results showed similar trends, with direct translation scoring 19.07, phrase-level pivoting 21.02, and sentence-level pivoting 20.33. Ahmadnia et al. further proposed a hybrid model combining direct and triangulated pivoting, which enhanced BLEU scores to 21.88 for System(1) and 22.02 for System(2), demonstrating the effectiveness of merging pivot-based and direct translation models.

Habash and Hu [7] also investigated the use of English as a pivot language, focusing on Arabic-Chinese SMT. As in the study by Ahmadnia et al. [1], they examined the three techniques—direct translation, sentence-pivoting, and phrase-pivoting—on four databases of different sizes. Pivoting systems consistently outperformed direct translation, with the largest dataset (XL) yielding BLEU scores of 16.17 for direct translation, 16.88 for sentence-pivoting, and 17.29 for phrase-pivoting.

Collectively, these studies highlight the effectiveness of pivoting techniques in enhancing translation quality across various language pairs and models, especially with English as the pivot language. Notably, our research aims to explore pivot-based translation for Hebrew, a language with limited resources that, to the best of our knowledge, has not been extensively studied in the context of pivot language translation. By utilizing high resource language as a pivot language and examining its impact on translations into low-resource languages, we seek to address this gap and contribute to a deeper understanding of the potential for pivoting in underexplored languages.

3 BACKGROUND

MT has evolved alongside technological advancements, becoming increasingly vital for global communication needs. This automated translation between languages [22] includes: (1) Statistical MT (SMT) systems, which, based on statistical probabilities, are capable of producing high-quality translations only if large bilingual corpora are available [25]. (2) Neural MT (NMT), which is based on deep learning techniques and is the dominant method in the field of machine learning [22]. Today, recent advances in MT have introduced diverse approaches for translating between language pairs.

Large Language Models (LLMs) like LLaMA represent a significant leap in the evolution of machine translation. LLaMA is an advanced collection of foundation language models, ranging from 7 billion to 65 billion parameters, designed to deliver state-of-the-art performance using open and publicly available datasets. Inspired by the Chinchilla scaling laws, LLaMA demonstrates that optimal results are achieved not merely through larger model sizes but by leveraging smaller models trained on extensive data within a compute budget. These models are built upon the transformer architecture and incorporate innovations like input normalization within transformer sub-layers, rotary positional embeddings instead of absolute positional embeddings, and the SwiGLU activation function. The models are trained using the AdamW optimizer, showcasing an efficient approach to constructing large-scale, high-performing language models. Although these models excel in multilingual understanding, they are particularly effective for high-resource languages while presenting opportunities for further enhancement in low-resource contexts [20].

Google Translate[1], which initially employed statistical methods, transitioned to neural machine translation in 2016, significantly

[1] https://en.wikipedia.org/wiki/Google_Translate

improving accuracy and fluency. Using the transformer architecture, a neural network design optimized for language processing, Google Translate processes vast parallel corpora to achieve effective zero-shot translation capabilities across multiple language pairs, even for low-resource languages. Its strength lies in its extensive language coverage and integration with Google's infrastructure, enabling rapid and adaptable translations for various contexts[2].

NLLB-200 (No Language Left Behind), developed by Meta, is a state-of-the-art neural machine translation model supporting over 200 languages, with a special focus on low-resource languages. It incorporates LASER-3 embeddings, which are multilingual embeddings trained to represent multiple language families, along with a sparsely gated mixture-of-experts architecture that enhances scalability across languages [4]. To improve translation quality, NLLB-200 applies self-supervised learning on monolingual data, boosting generalization and reducing overfitting. Trained on diverse datasets, its multilingual architecture excels in scenarios with scarce parallel corpora. NLLB-200's focus on inclusivity ensures high-quality translations for historically underserved languages[3].By incorporating NLLB-200 as one of our models, we can evaluate whether using a pivot language for translation remains effective when applied with a model specifically tailored for low-resource language translation [4].

OPUS-MT, including models developed by Helsinki-NLP group, is an open-source initiative offering over 1,000 pre-trained models for bilingual and multilingual translation. It leverages the Marian-NMT framework for efficient training and decoding, utilizing techniques like fine-tuning and backtranslation to enhance domain-specific performance [19]. The models employ bidirectional encoder representations from transformers (BERT), which analyze text bidirectionally for deeper contextual understanding [6]. In this study, we utilize OPUS-MT models fine-tuned for specific language pairs, including Hebrew-English, English-Finnish, and English-Ukrainian for pivot-based translation, as well as Hebrew-Finnish and Hebrew-Ukrainian for direct translation.

Despite progress, MT still faces challenges, particularly with low-resource languages, which lack extensive parallel corpora to train models effectively [22, 25]. To address this challenge, Wu and Wang [25] introduced a translation model called **Pivot Model** (also known as indirect translation). Their model is based on phrase-based SMT and uses a high-resource language as a Pivot Language to bridge between low-resource language pairs.

To translate between languages L_s and L_t, the pivot language L_p is used, for which large bilingual corpora exist for the pairs L_s-L_p and L_p-L_t. Two models are trained using these corpora, one for each language pair, and a translation model for L_s-L_t is derived from them, known as the Pivot Model.

Translation quality is commonly evaluated using automated metrics that compare machine-generated translations to reference texts. One of the most widely used metrics, Bilingual Evaluation Understudy (BLEU), measures translation quality by calculating n-gram overlaps between candidate and reference translations. BLEU incorporates a brevity penalty to address length mismatches, ensuring a balance between precision and fluency. However, BLEU has limitations in capturing semantic meaning or alternative phrasing. For example, BLEU might penalize a translation for not matching word-for-word even if the meaning is correct, which makes it less nuanced compared to other metrics like METEOR or COMET [13].

Metric for Evaluation of Translation with Explicit Ordering (METEOR) enhances BLEU by introducing features such as stemming, synonym matching, and a weighted harmonic mean of precision and recall. These features allow METEOR to be more sensitive to semantic similarities and alternative expressions. For instance, in a Hebrew-Finnish translation, METEOR would recognize that "Rauha sinulle" ("Peace to you") and "Rauha teille" ("Peace to you all") convey the same meaning, even though the words are different. By considering synonyms and variations in word forms, METEOR rewards these translations as semantically equivalent, providing a more nuanced measure of translation quality than BLEU, which only rewards exact word matches [2].

Crosslingual Optimized Metric for Evaluation of Translation (COMET), a more recent metric, leverages multilingual embeddings and neural networks trained on human-annotated datasets to evaluate translations. It excels in capturing both semantic adequacy and fluency, aligning closely with human judgments and outperforming traditional metrics. For example, if the Hebrew sentence "שלום עליכם" is translated into Finnish as "Rauha teille" (correct) and "Rauha maailma" ("Peace world"), COMET would assign a higher score to the former due to its semantic alignment and fluency. COMET has demonstrated a stronger correlation with human judgments compared to traditional metrics, making it a valuable tool for translation evaluation [16]. COMET has three basic architectures and many versions that were published during the years. The basic architectures include the estimator model which uses the following vector and feeds it to feed forward network for regression:

$$x = [h; r; h \odot s; h \odot r; |h - s|; |h - r|]$$

where h is the translation being evaluated, s is the source sentence, and r is the reference. The latest version COMET22 [15], which use ensemble of the estimator model with a score from a sequence tagger model that gives a binary tag for each word in the translation. The tags are combined to one score according to the following equation:

$$\hat{y}_{\text{tags}} = 1 - w \times \frac{\sum_i^{N_S} \mathbb{1}[S_i = \text{BAD}]}{N_S}$$

where S_i is the i-token in the sequence, N_S is the number of tokens in the sequence, and w is a severity penalty for BAD tags.

In summary, while MT has made significant progress through the development of various models, challenges remain, particularly for low-resource languages. Pivot-based approaches, offer promising solutions by leveraging high-resource languages as intermediaries for translation.

4 METHODOLOGY

4.1 Overview

This study explores the use of pivot-based translation to improve MT quality from Hebrew to low-resource languages. The research focuses on using English as an intermediary pivot language to evaluate its impact on enhancing translations into Finnish and Ukrainian.

[2]https://research.google/blog/unlocking-zero-resource-machine-translation-to-support-new-languages-in-google-translate/

[3]https://ai.meta.com/blog/nllb-200-high-quality-machine-translation/

To evaluate the effectiveness of different translation models, we will compare the pre-trained multilingual model NLLB-200 and the pair-specific Helsinki-NLP model. Additionally, we will assess Google Translate and Llama 3.1. The evaluation involves both direct translation from Hebrew to Finnish (H-F) and from Hebrew to Ukrainian (H-U) and pivot-based translation, where Hebrew is first translated into English before being translated into the target languages.

4.2 Triangulation Translation

The term triangulation translation refers to the process of using a **pivot language** as an intermediary between the source and target languages. Triangulation translation involves three languages: the source language (L_s), pivot language (L_p), and target language (L_t). The source language is first translated into the pivot language, and then from the pivot language into the target language. This intermediary step enables translation between language pairs that lack direct parallel corpora.

Parallel corpora, which are essential for training high-quality translation models, are not readily available for many language pairs, especially for less commonly spoken languages [3]. Triangulation translation addresses this gap by leveraging existing bilingual corpora between the source and pivot language (L_s-L_p) and between the pivot and target language (L_p-L_t) to create a robust translation model. English is often selected as the pivot due to its prevalence in multilingual corpora, although other languages can serve as pivots depending on available resources [21].

This method is particularly helpful when direct bilingual corpora are lacking, but sufficient corpora exist for the pivot language in both directions[3]. Figure 1 illustrates the steps of the pivot model.

Figure 1: The pivot model

Pivot translation can be implemented by using the pivot language as an intermediary in various translation methods, with the most common translation approaches being phrase-level and sentence-level translation.

4.3 Translation methods

4.3.1 *Phrase-level translation:*

Phrase-level translation involves segmenting sentences into smaller units, known as phrases, which are translated individually and then recombined to form the final output. This approach ensures that the translated phrases are properly aligned and adjusted to fit the grammatical structure of the target language, producing a coherent and fluent sentence [10].

Figure 2 demonstrates the use of a pivot language in a model that performs phrase-level translation. In the first stage, the model translates from the source language to the pivot language, operating behind the scenes at the phrase level. The sentence is divided into segments, each segment is translated independently, and the segments are then recombined into a full sentence. This process is repeated in the second stage, translating from the pivot language to the target language.

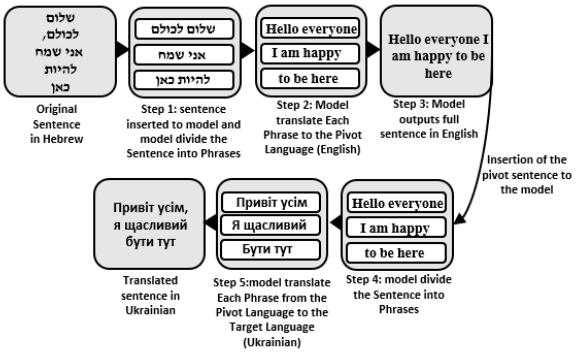

Figure 2: Phrase-Level translation model with Pivot Language Integration

4.3.2 *Sentence-level translation:*

In sentence-level translation, the process involves directly translating the source sentence into the target language in a single step, resulting in the final output [21]. Figure 3 illustrates the use of a pivot language in our experiment, incorporating a model that performs sentence-level translation.

Figure 3: Sentence-Level translation model with Pivot Language Integration

All the models we will use in the study work on the basis of sentence-level translation, with Helsinki-NLP, as an exception, as it uses phrase-level translation in some cases.

4.4 Novel Contributions

This research tackles the broader challenge of Hebrew MT, a language that remains significantly underexplored in the field of natural language processing. Despite Hebrew's linguistic richness and global significance, its translation to other languages, particularly low-resource ones like Finnish and Ukrainian, has received limited attention. To address this gap, we introduce and evaluate pivot-based translation strategies for H-F and H-U pairs, using these pairs as a testbed for method evaluation.

Our study compares between direct and pivot-based translation methods across state-of-the-art MT model and employs advanced evaluation metrics to comprehensively assess translation accuracy, fluency, and semantic adequacy. This work not only demonstrates

the potential of pivot-based methods for improving Hebrew translation but also contributes a structured framework for evaluating translation strategies in low-resource language settings, paving the way for further exploration of Hebrew and similarly underrepresented languages.

5 EMPIRICAL EVALUATION

5.1 Research Questions

This study aims to answer the following research questions:

1. Does pivot-based translation yield higher translation quality than direct translation across Hebrew-Finnish and Hebrew-Ukrainian language pairs?
2. Which translation models achieve the best overall performance for each method (direct and pivot-based translations) and the language pairs based on the selected evaluation metrics?

5.2 Research Hypothesis

This study tests the hypothesis that pivot-based translation, which utilizes the abundant high-resource language data available for English, will outperform direct translation in enhancing the quality of translations from Hebrew into low-resource languages like Finnish and Ukrainian.

5.3 Experimental Setup

5.3.1 *Datasets.*

The data sets used in this research are sourced from the TED 2020 subset of the OPUS corpus[4], a widely recognized resource for multilingual translation studies. Specifically, the H-F dataset contains 44,057 sentences, while the H-U dataset includes 190,042 sentences, providing a substantial basis for evaluating varying translation strategies.

Upon reviewing the dataset, we identified several translation mismatches and inconsistencies in the H-U and H-F sentence pairs. These issues underscored the need for a thorough cleaning process to preserve the integrity of the data.

To address this, we utilized the multilingual-e5-large sentence embedding model by Intfloat [23] to generate vector embeddings for each sentence in the pairs. Cosine similarity was then calculated between these embeddings to detect inconsistencies.

The Figures 4 and 5 present the distribution of cosine similarities for the H-U and H-F language pairs respectively, which served as a basis for identifying and removing unreliable translations from the dataset.

As shown in Figures 4 and 5, for both language pairs, the majority of the sentence pairs exhibit a similarity score of at least 0.85, with relatively few scoring 0.75 or lower. However, sentences with a similarity score of 0.75 or below are the most likely to represent incorrect translations between the languages. For instance, the pair 'כי הפעם הצלחתי לענות ממש במהירות.' and 'Olin' has a similarity score of 0.737. Upon review, the Hebrew sentence translates to "This time I successfully answered really fast," while the Finnish sentence translates to "I was," showing a complete mismatch. As a result, all sentence pairs with a similarity score below 0.75 were filtered out.

[4]https://opus.nlpl.eu/TED2020/corpus/version/TED2020

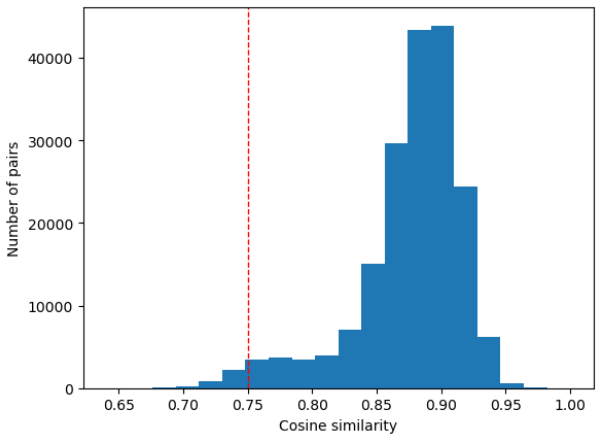

Figure 4: Histogram of the cosine similarity between semantic embeddings of Hebrew and Ukrainian sentences

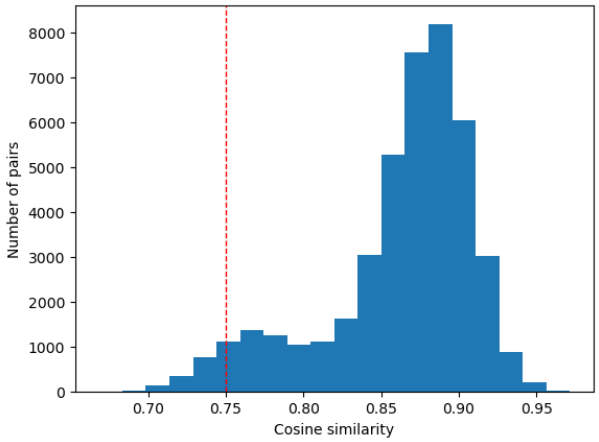

Figure 5: Histogram of the cosine similarity between semantic embeddings of Hebrew and Finnish sentences

Additionally, we excluded sentences containing English words, as their presence could impact both the evaluation metrics and the models' performance. After this filtering process, we selected 20,000 high-quality sentence pairs that are identical in both the Ukrainian and Finnish datasets. This ensures consistency across the two languages and provides a balanced and reliable dataset for further analysis.

5.3.2 *Experiment Pipeline.*

This experiment evaluates the impact of using a pivot language (English) on translation quality for H-F and H-U language pairs. The process involves a two-step approach for each language pair, employing and compering the models NLLB-200, Helsinki-NLP, Google Translate, and Llama, with results assessed using three key metrics: BLEU, METEOR, and COMET.

In the first stage, direct translations are performed from Hebrew to the target languages (Finnish or Ukrainian) using each of the four models. The translated outputs are evaluated using the three metrics, establishing a baseline measure of translation quality for each model under the direct translation approach.

The second stage introduces English as a pivot language. In this approach, Hebrew source texts are first translated into English, and the English translations are subsequently translated into the target language (Finnish or Ukrainian), with the same model used for both steps. All four models are evaluated to assess the effects of incorporating a pivot language into the translation process and to explore potential variations in performance under this method.

The results are then analyzed to determine which approach yields better performance for each model. The analysis also examines the strengths and weaknesses of each model. Since each metric prioritizes different aspects of translation quality, some models may excel in producing precise, word-for-word translations, while others may be more adept at generating translations that resemble human-like fluency and naturalness. This exploration sheds light on how each model handles various dimensions of translation quality and whether performance improves when an intermediate language is introduced. The workflow of the pipeline is presented in Figure 6.

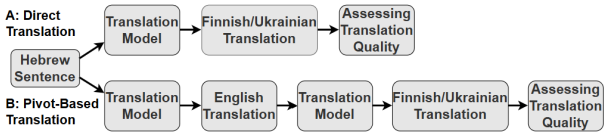

Figure 6: Overview of the experimental setup comparing direct (A) and pivot-based (B) translation approaches for Hebrew-Finnish and Hebrew-Ukrainian language pairs

5.4 Statistical Tests

To evaluate the statistical significance of the results, we performed a paired t-test at the sentence level to determine whether the mean differences between the two groups—Direct and Pivot-Based translation methods—were statistically significant [12]. This analysis was conducted for each of the translators (NLLB-200, Helsinki-NLP, Google Translate, and Llama) across the three metrics: BLEU, METEOR, and COMET. By performing the paired t-test at the sentence level, we captured variations at a more granular level, reflecting differences in translation performance for individual sentences. The sign of the t-statistic (positive or negative) indicates the direction of the difference between the two groups. Specifically, a positive t-statistic implies that, on average, the values in the first group (e.g., direct translation) are greater than those in the second group (e.g., pivot translation), indicating that the direct translation performed better. Conversely, a negative t-statistic would suggest the opposite. Furthermore, by using a paired t-test, we accounted for the fact that both methods were evaluated on the same dataset, ensuring that observed variations were not influenced by differences in test samples. A p-value threshold of 0.05 was applied to identify significant differences in performance.

5.4.1 *Ablation Study Design.*

The ablation study is designed to isolate the impact of key components in the translation process. It includes the following steps:

Direct Translation (Baseline): For each sentence, a reference translation is provided in the dataset. This baseline translation serves as the standard for comparison, against which the translations generated by the models, whether directly from Hebrew to the target language or via the pivot language (English), are evaluated.

Pivot Language (English): Evaluating the translation quality when English is used as an intermediary language between Hebrew and the target language.

Model-Specific Ablations: Comparing how each model (NLLB-200, Helsinki-NLP, Google Translate, Llama) performs with and without the pivot language to understand its impact.

Metric Comparison: Analyzing the results across three metrics (BLEU, METEOR, COMET) to analyze and evaluate each model based on different aspects of translation quality, such as precision, fluency, and semantic accuracy.

This design allows us to determine the contribution of the pivot language and model choice to the overall translation quality.

5.5 Experimental Settings

All experiments were conducted on virtual machines running the CentOS Linux 7 (Core) operating system, equipped with 30 GB of memory and an NVIDIA GeForce RTX 3090 GPU. This computational setup was chosen to efficiently manage large-scale translation tasks, ensuring both consistency and reproducibility of results for models such as NLLB-200, Helsinki-NLP, and Llama.

All scripts were implemented in Python 3.10, utilizing the following libraries and frameworks:

- ***Hugging Face Transformers***: For running experiments with the Helsinki-NLP, NLLB-200 and Llama models, specifically using the *pipeline* module for easy integration and inference.
- ***google-cloud***: For accessing the Google Translate API for translation tasks.
- ***sacrebleu***: For evaluating translations using the BLEU metric.
- ***evaluate***: For evaluating translations using the METEOR metric.
- ***comet-ml***: For evaluating translations using the COMET metric.
- ***sentence-transformers***: For generating sentence embeddings using the *intfloat/multilingual-e5-large* pre-trained model. This was used for computing cosine similarity to filter out unreliable translation pairs.
- ***sklearn***: Specifically the `cosine_similarity` function from *sklearn.metrics.pairwise*, which was used to compute the similarity between translation sentence pairs.
- ***xml.etree.ElementTree***: For extracting translation pairs from XML files.

This setup ensures standardized, efficient, and reproducible execution of all experiments.

5.6 Models Configurations

- **Open-Source Models:** The open-source models NLLB-200 and Helsinki-NLP were utilized in their pre-trained, "off-the-shelf" versions. The maximum number of new tokens was set to **600** for generating translations.

- **Google Translate API:** The Google Translate API was employed with its **default parameters**, providing a standard baseline for translation performance.
- **Llama 3.1 (8B):**
  - **Maximum Tokens:** The maximum number of new tokens was set to **600**.
  - **Temperature:** A value of **0.3** was used to control the randomness of the outputs, favoring more deterministic translations.
  - **Top-p:** Set to **0.9**, ensuring that only the most probable tokens contributed to the output, improving coherence.
  - **Prompt Settings:**
    System Prompt: "You are a professional translator. Always translate exactly to 'target lang' language only and return only the translation."
    User Prompt: "Translate this 'source lang' text to 'target lang': 'input text'."

5.7 Results

The results for H-F translation are presented in Table 1, while those for H-U translation are shown in Table 3. Both tables compare the performance of direct and pivot-based translation (via English) across BLEU, METEOR, and COMET metrics, with reported values at the corpus level.

To evaluate the impact of pivoting on translation quality, paired t-tests were conducted for each model across all metrics, comparing direct translation with the pivot-based method. The t-test results for H-F and H-U are summarized in Tables 2 and 4, respectively.

5.7.1 *Results for Hebrew-Finnish Translation.*

For *direct translation*, Google Translate achieved the best performance across all metrics, with a BLEU score of 11.5, a METEOR score of 0.42, and a COMET score of 0.85. These results highlight the capabilities of Google Translate in direct translation, even for low-resource language pairs such as Hebrew-Finnish. Google Translate's dominance in both BLEU and COMET underscores its ability to balance lexical accuracy and fluency, likely due to its extensive training on large, diverse datasets. The Helsinki model followed closely, with slightly lower scores, particularly in BLEU (10.44) and METEOR (0.36). The NLLB-200 and Llama models demonstrated comparatively lower performance, with Llama achieving the lowest scores overall. The relatively low performance of NLLB-200 may be attributed to its broader training on multiple languages, which may not have fully captured the intricacies of the H-F language pair. Llama, being a more general-purpose model, may not have been fine-tuned for the specific nuances of these languages, leading to its lower scores.

In *pivot-based translation*, the results indicate an overall improvement for some models, with slight improvements observed in METEOR and COMET scores across all models. However, BLEU scores decreased for the Helsinki (10.44 vs. 9.93) and Google Translate (11.5 vs. 7.8) models, while NLLB-200 and Llama showed small increases. This suggests that while pivoting may help in preserving the semantic meaning and fluency, it can lead to a decrease in lexical accuracy for certain models. The decrease in BLEU for Google Translate could be due to the loss of direct lexical mapping when using a pivot language, while Helsinki experienced a more moderate decrease.

Paired t-test results (Table 2) showed significant differences between direct and pivot-based methods. NLLB-200 direct translation outperformed pivoting in BLEU and METEOR, as reflected by its positive t-statistics (BLEU: 2.858, METEOR: 4.906). In contrast, Helsinki, Google Translate, and Llama showed negative t-statistics across all metrics, indicating that pivoting improved fluency and semantic accuracy, despite a trade-off in lexical precision for some models.

Translation Method	Model	Metrics		
		BLEU	METEOR	COMET
Direct Translation	NLLB-200	6.14	0.33	0.79
	Helsinki	10.44	0.36	0.82
	Google Translate	**11.5**	**0.42**	**0.85**
	Llama	5.42	0.29	0.77
Pivot-Based Translation	NLLB-200	8.56	0.33	0.8
	Helsinki	**9.93**	0.39	0.84
	Google Translate	7.8	**0.42**	**0.86**
	Llama	6.0	0.3	0.79

Table 1: The evaluation results for the Hebrew-Finnish language pair

Metric	Model	T-statistic (P-value)
BLEU	NLLB-200	2.858 (0.004*)
	Helsinki	-9.2 (3.93×10^{-20}*)
	Google Translate	-9.75 (2.15×10^{-22}*)
	Llama	-15.28 (1.98×10^{-52}*)
METEOR	NLLB-200	4.906 (9.36×10^{-7}*)
	Helsinki	-20.76 (9.79×10^{-95}*)
	Google Translate	-9.33 (1.07×10^{-20}*)
	Llama	-8.99 (2.46×10^{-19}*)
COMET	NLLB-200	-10.87 (1.89×10^{-27}*)
	Helsinki	-24.68 (1.61×10^{-132}*)
	Google Translate	-17.85 (1.003×10^{-70}*)
	Llama	-15.26 (2.67×10^{-52}*)

Table 2: Paired t-test Results for Direct vs. Pivot-Based Translation Methods in the Hebrew-Finnish Language Pair
*Note: * indicates p-value < 0.05, representing statistically significant differences.*

5.7.2 *Results for Hebrew-Ukrainian Translation.*

For Hebrew-Ukrainian translation, *pivot translation* generally outperformed *direct translation* across most models, particularly in terms of BLEU and COMET scores. The Helsinki model demonstrated a significant improvement in both BLEU (from 15.63 to 25.83) and COMET (from 0.77 to 0.8) with pivoting. This trend aligns with the Hebrew-Finnish results, where Helsinki similarly benefited from pivoting. In both cases, pivoting likely improved Helsinki's ability to handle semantic alignment, as the intermediary

language aids in disambiguating meanings and enhancing overall fluency.

In contrast, Google Translate performed better in BLEU with direct translation (24.36 vs. 13.34), consistent with the Hebrew-Finnish results (11.5 vs. 7.8). This suggests that Google Translate excels in lexical accuracy with direct translation. However, pivoting improved COMET and METEOR scores, demonstrating the trade-off between lexical precision (better in direct translation) and semantic fluency (improved with pivoting).

The NLLB-200 model showed a BLEU increase with pivoting in both language pairs (from 12.5 to 17.19 in H-U and from 6.14 to 8.56 in H-F), indicating that the pivot language provides valuable contextual support, improving lexical accuracy for this model. This improvement suggests that pivoting helps NLLB-200 disambiguate translations, resulting in better alignment with reference translations.

Llama showed modest improvements in BLEU, METEOR, and COMET when pivoting, suggesting that, as a general-purpose model, it benefits less from pivoting than specialized models like Helsinki or NLLB-200. This mirrors the Hebrew-Finnish results, where Llama also showed only slight improvements with pivoting.

The paired t-test results reveal statistically significant differences between direct and pivot-based translation methods for most models and metrics, supporting the trends seen in the evaluation scores. For NLLB-200, no significant differences were found in BLEU and METEOR, indicating that its performance remained stable between the two translation methods.

Translation Method	Model	Metrics		
		BLEU	METEOR	COMET
Direct Translation	NLLB-200	12.5	0.37	0.78
	Helsinki	15.63	0.39	0.77
	Google Translate	**24.36**	**0.42**	**0.84**
	Llama	10.87	0.32	0.79
Pivot-Based Translation	NLLB-200	17.19	0.37	0.79
	Helsinki	**25.83**	0.38	0.8
	Google Translate	13.34	**0.42**	**0.85**
	Llama	18.72	0.33	0.79

Table 3: The evaluation results for the Hebrew-Ukrainian language pair

5.8 Discussion

The study's results revealed both expected and unexpected trends when comparing direct and pivot-based translation methods. Our initial expectations were that all models and metrics would show clear improvements in translation quality using the pivot-based translation method. However, a notable exception emerged with the BLEU metric for Google Translate. In both language pairs, H-F and H-U, direct translation outperformed pivot-based translation in corpus-level BLEU scores. BLEU, as discussed in the background section, is highly sensitive to exact n-gram matches, making it particularly vulnerable to changes in word order. The pivot-based translation process, which involves a two-step translation through an intermediary language, often preserved sentence meaning but introduced structural shifts, negatively impacting BLEU scores. Google Translate relies on large, openly edited, and crowdsourced datasets, which provide broad language coverage but can introduce inconsistencies in translations[5]. While crowdsourced data can enhance language adaptability, it also leads to variations in quality, especially with widely spoken languages like English, which has a large pool of contributors. This variability in the intermediary language can cause structural misalignments in the translation, further impacting BLEU scores, which prioritize lexical and structural fidelity.

Metric	Model	T-statistic (P-value)
BLEU	NLLB-200	0.09 (0.92)
	Helsinki	3.78 (0.0001*)
	Google Translate	-9.07 (1.19×10^{-19}*)
	Llama	-7.54 (4.69×10^{-14}*)
METEOR	NLLB-200	1.84 (0.064)
	Helsinki	5.36 (8.18×10^{-8}*)
	Google Translate	-5.8 (6.56×10^{-9}*)
	Llama	-5.32 (1.04×10^{-7}*)
COMET	NLLB-200	-17.43 (1.5×10^{-67}*)
	Helsinki	-44.23 (0.0*)
	Google Translate	-18.39 (6.44×10^{-75}*)
	Llama	-8.15 (3.61×10^{-16}*)

Table 4: Paired t-test Results for Direct vs. Pivot-Based Translation Methods in the Hebrew-Ukrainian Language Pair
*Note: * indicates p-value < 0.05, representing statistically significant differences.*

In contrast, metrics like METEOR and COMET, which reward semantic preservation and accommodate structural variations, demonstrated either maintained or improved results with the pivot-based approach.

For instance, Table 5 highlights a Hebrew-to-Finnish sentence example from the Google Translate model where the direct translation achieved a higher BLEU score of 41.63, demonstrating closer alignment with the reference sentence's n-gram structure compared to the pivot translation. The pivot translation scored lower at 12.95 due to structural adjustments and word substitutions, although it successfully retained the sentence's semantic meaning. Unlike BLEU, METEOR and COMET reflected the pivot translation's ability to preserve meaning, assigning comparable or slightly higher scores.

Across all models and language pairs, COMET consistently favored pivot-based translation, as indicated by negative t-statistics and slight score improvements. This reflects COMET's emphasis on semantic adequacy and fluency over structural fidelity. Models like NLLB-200 and LLaMA, which leverage multilingual and self-supervised learning, benefited from pivoting, while Helsinki-NLP struggled, showing lower BLEU scores. Since BLEU prioritizes exact n-gram matches, the transformations introduced by pivoting reduced lexical precision, particularly for models tailored to specific language pairs like Helsinki-NLP.

[5]https://support.google.com/translate/thread/6447871/wrong-translation-but-flagged-as-verified-by-translate-community?hl=en

Hebrew	אני רוצה שתדעו שאני, אחי ואחותי באמת אוהבים לאכול צ'יפס כרוב אפוי.
Finnish Reference	**Haluan** teidän **tietävän, että minä, veljeni ja** sisareni itse asiassa pidämme paistetuista kaalilastuista.
H-F	**Haluan** sinun **tietävän, että minä, veljeni ja** siskoni syömme todella mielellään uunikaalialastuja.
H-English-F	**Haluan** sinun **tietävän, että** veljeni, sisareni **ja** minä todella tykkäämme syödä uunikaalialastuja.

Table 5: An example of Hebrew to Finnish translation by Google Translate. The bolded words indicate the similar words between the reference sentence and the translations that appear in the same positions.

One key limitation of this study is the choice of English as the sole pivot language. English was selected due to its high corpus availability and extensive use in machine translation, ensuring accessibility across Hebrew, Finnish, and Ukrainian. However, despite these advantages, English does not effectively bridge the linguistic differences between these languages, as they belong to distinct language families with unique grammatical structures and vocabularies. This mismatch may introduce inaccuracies and lead to the loss of cultural and contextual nuances in translation. Future research could explore alternative pivot languages that share greater linguistic or structural similarities with either the source or target language, potentially improving translation quality and providing deeper insights into the impact of pivot language selection [14].

Lastly, the choice of BLEU, METEOR, and COMET as evaluation metrics offered complementary perspectives on translation quality. BLEU is widely used for its simplicity and efficiency, though its reliance on exact matches can limit its applicability for translations with structural variations. METEOR improves upon BLEU by considering synonyms and word forms, while COMET excels in assessing semantic adequacy and fluency through its neural framework. These metrics proved well-suited for the study. BLEURT [17], while a promising evaluation metric, was not included due to its optimization for English and limited generalizability to non-English language pairs like Finnish and Ukrainian without additional fine-tuning.

In conclusion, while pivot-based translation methods show promise in enhancing semantic accuracy and fluency, the results highlight the importance of model-specific considerations and the choice of evaluation metrics for improving translation performance.

6 CONCLUSIONS AND FUTURE WORK

In this study, we investigated the use of a pivot language to improve translation quality between low-resource language pairs: Hebrew-Finnish and Hebrew-Ukrainian. Our findings suggest that pivot-based translation, utilizing English as an intermediary, generally outperforms direct translation in terms of translation quality across both language pairs. Statistical t-tests confirmed that while pivot-based translation tends to perform better overall, certain models showed better results with direct translation. This highlights the importance of selecting the appropriate translation method and evaluation metric based on the specific goals of the translation task.

The primary limitation of this study include the reliance on English as the sole pivot language, which may have limited the effectiveness of translations in this study. In future investigations, exploring alternative pivot languages with closer linguistic ties to Hebrew, Finnish, or Ukrainian could offer significant improvements, especially if these languages have high-resource status or large bilingual corpora [14]. For example, Russian, an East Slavic language, shares substantial grammatical and lexical similarities with Ukrainian [6], making it a strong candidate for translations involving Ukrainian, particularly if large Russian-Ukrainian corpora are available. Similarly, Hungarian, a Finno-Ugric language like Finnish, shares structural features such as agglutinative morphology and an extensive case system [7], which could enhance translation accuracy for Finnish-related pairs, provided there are sufficient bilingual resources. Arabic, as a Semitic language closely related to Hebrew, shares root-based morphology and similar verb structures [8], offering a natural linguistic bridge for Hebrew translations. Leveraging such linguistically similar pivot languages with rich corpora has the potential to reduce errors, improve semantic preservation, and enhance fluency in machine translation systems.

Additionally, future work could explore dynamic pivot language selection, where translations are generated through multiple pivot languages, with the system integrating the best elements from each output [5]. For example, translating Hebrew to Ukrainian could involve parallel translations through Russian and Belarusian, dynamically selecting the path that achieves the highest score based on evaluation metrics. Furthermore, investigating the use of multiple pivot languages sequentially, such as Hebrew $\rightarrow$ Arabic $\rightarrow$ Russian $\rightarrow$ Ukrainian, could help mitigate errors introduced in single-pivot systems and improve translation consistency across complex linguistic paths [11].

Another promising direction is the exploration of hybrid translation methods, which combine direct and pivot-based approaches [1]. In this approach, translations would be generated both directly and via a pivot language, and the system would select the translation with the highest score. By dynamically choosing the best-performing translation, this method ensures improved outcomes tailored to the specific priorities of the translation task.

The exploration of more advanced translation techniques, such as GPT-based models [9] and emerging methods like cross-lingual transfer learning or zero-shot translation, could enhance pivot-based translation approaches. Expanding the range of techniques tested may lead to better strategies for bridging linguistic gaps and preserving meaning in low-resource language pairs.

Overall, this study demonstrated the potential of pivot-based translation for improving translation quality between Hebrew-Finnish and Hebrew-Ukrainian. The findings highlight the value of pivot language methods, but further exploration is needed to fully harness their potential. With continued experimentation and innovation, the field of pivot-based translation holds great promise for bridging linguistic gaps and improving machine translation quality.

[6] https://en.wikipedia.org/wiki/Old_East_Slavic?utm
[7] https://en.wikipedia.org/wiki/Uralic_languages?utm
[8] https://en.wikipedia.org/wiki/Semitic_languages

REFERENCES

[1] Benyamin Ahmadnia, Javier Serrano, and Gholamreza Haffari. 2017. Persian–Spanish Low-Resource Statistical Machine Translation Through English as Pivot Language. arXiv:1709.03411 [cs.CL]

[2] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 65–72.

[3] Trevor Cohn and Mirella Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. 728–735.

[4] Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672* (2022).

[5] Raj Dabre, Aizhan Imankulova, Masahiro Kaneko, and Abhisek Chakrabarty. 2021. Simultaneous multi-pivot neural machine translation. *arXiv preprint arXiv:2104.07410* (2021).

[6] Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[7] Nizar Habash and Jun Hu. 2009. Improving Arabic-Chinese Statistical Machine Translation using English as Pivot Language. , 173–181 pages.

[8] Kenji "Imamura, Masao Utiyama, editor: "Utiyama Masao Sumita, Eiichiro", and Rui" Wang. 2023. "Pivot Translation for Zero-resource Language Pairs Based on a Multilingual Pretrained Model". In *"Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track"*. 348–359.

[9] Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is ChatGPT a good translator? A preliminary study. *arXiv preprint arXiv:2301.08745* 1, 10 (2023).

[10] Philipp Koehn. 2009. *Phrase-Based Models*. Cambridge University Press, 127–154.

[11] Shivam Mhaskar and Pushpak Bhattacharyya. 2022. Multiple Pivot Languages and Strategic Decoder Initialization Helps Neural Machine Translation. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*. 9–14.

[12] Prabhaker Mishra, Uttam Singh, Chandra M Pandey, Priyadarshni Mishra, and Gaurav Pandey. 2019. Application of student's t-test, analysis of variance, and covariance. *Annals of cardiac anaesthesia* 22, 4 (2019), 407–411.

[13] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[14] Michael Paul, Hirofumi Yamamoto, Eiichiro Sumita, and Satoshi Nakamura. 2009. On the importance of pivot language selection for statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. 221–224.

[15] Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 578–585.

[16] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. *arXiv preprint arXiv:2009.09025* (2020).

[17] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. BLEURT: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696* (2020).

[18] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning. arXiv:2008.00401 [cs.CL]

[19] Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT–building open translation services for the world. In *Proceedings of the 22nd annual conference of the European Association for Machine Translation*. 479–480.

[20] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[21] Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. 484–491.

[22] Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2022. Progress in Machine Translation. *Engineering* 18 (2022), 143–153.

[23] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report. arXiv:2402.05672 [cs.CL] https://arxiv.org/abs/2402.05672

[24] Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending Multilingual BERT to Low-Resource Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 2649–2656.

[25] Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. *Machine Translation* 21 (2007), 165–181.

[26] Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. 2023. Plug: Leveraging pivot language in cross-lingual instruction tuning. *arXiv preprint arXiv:2311.08711* (2023).