Machine Translation of Texts from Languages with Low Digital Resources: A Systematic Review

Hermilo Benito-Santiago[1(✉)], Diana Margarita Córdova-Esparza[1], Noé Alejandro Castro-Sánchez[2], and Ana-Marcela Herrera-Navarro[1]

[1] Facultad de Informática, Universidad Autónoma de Querétaro, Queretaro, Mexico
hsantiago13@alumnos.uaq.mx, {diana.cordova,mherrera}@uaq.mx
[2] Centro Nacional de Investigación y Desarrollo Tecnológico, Cuernavaca, Mexico
noe.cs@cenidet.tecnm.mx

Abstract. This research conducted a systematic review of related works on machine translation of languages with low digital resources. First, we carried out the information search in the databases: ScienceDirect, IEEE Xplore, ACM Digital Library. Eighteen articles were collected following inclusion and exclusion criteria, considering a search period from 2016 to 2022. Subsequently, we analyzed and classified these articles according to the libraries developed and/or used based on machine learning, statistics, or grammar. The results indicate that pretraining and morphological segmentation techniques with finite state machines and machine learning techniques improve the translation of languages with low digital resources. In addition, according to the articles compiled in the specialized databases, in Mexico, unlike other countries that we analyzed, there are few publications on the translation of languages with low digital resources, and we mostly found research papers published in international conferences.

Keywords: Machine translation · Parallel corpus · Languages with low digital resources

1 Introduction

Languages with low digital resources are those languages that do not have a large amount of written or digitized documentation. It may also be the case that this documentation exists but is not published. The reduced number of language speakers means no documentation is generated about it. These languages represent a significant challenge in Artificial Intelligence knowledge, particularly in natural language processing, for two main reasons: the first refers to the scarcity of corpus since, to carry out the experimentation, a set of data is required. With a considerable size that can be processed to obtain results; the second refers to rethinking and adapting existing methods that have been used with languages with characteristics different from languages that are considered to have low digital resources [1].

O. Pichardo Lagunas et al. (Eds.): MICAI 2022, LNAI 13613, pp. 41–56, 2022.
https://doi.org/10.1007/978-3-031-19496-2_4

According to Hedderich et al. [2] low-resource languages can be divided into:

– The availability of task-specific labels.
– Availability of untagged or domain-specific language.
– Availability of auxiliary data.

In this research work, we performed a literature review on the latest trends regarding machine translation methods for languages with limited digital resources. We considered specialized databases such as ScienceDirect, IEEE Xplore, ACM Digital Library, analyzing journal articles, book chapters, and proceedings.

This article aims to quantitatively and qualitatively analyze the state-of-the-art for automatic translation of languages with low digital resources. The quantitative analysis was carried out to have a perspective on the number of articles published by year, country, and by areas of knowledge. In contrast, the qualitative analysis categorizes articles with characteristics in common regarding the methodology they implement and the resources used.

This paper contains the following sections. Second we present, the method with the following phases, documentary search in databases, description of criteria used for select works, analysis, and categorization of works, and discussion. Third, we present conclusions and future work after reviewing works.

2 Method

In this work, we carried out a mixed systematic review, for which the information is analyzed quantitatively and qualitatively from the works that have been developed in the literature on machine translation of low-resource languages.

To carry out the systematic review, we adopted the phases of Palacios-Díaz and Escudero-Nahón [3]. Next, we detail each of the phases that make up this work's methodology.

2.1 Phase 1: Documentary Search

We carried out the documentary search in the databases: ScienceDirect, IEEE Xplore, and ACM Digital Library, selecting journal articles, book chapters, and conference proceedings. Words that include machine translation and languages of limited resources were taken into account due to the main focus of finding articles related to machine translation.

In the case of the term low-resources languages, we considered different ways that these languages can be named. According to the author Singh [4], they can be identified as less-studied languages, languages of scarce resources, less computerized languages , and less privileged languages. Other terms that Cieri mentions [5] are: low density, less taught, scarce resources, less resources, low resources, critical languages, and in danger of extinction.

Taking into account the terminologies mentioned above, we considered the following keywords: minority languages, fewer resources, languages in danger of extinction, language, and dialect; to find research with these characteristics. We replicated the following

query in each database: *machine translation AND (low resource OR scarce resource OR poor OR minoritized OR lesser resourced OR endangered) AND (dialect OR languages).*

According to the results obtained by entering the query string and following the selection criteria detailed below, we identified relevant articles related to machine translation of low-resource languages.

2.2 Phase 2: Description of Selection Criteria

In the searches in ScienceDirect, we obtained 1,592 results, in IEEE Xplore 259 results, in ACM Digital Library 395 results, in Proceedings (ACL Anthology) 3130 results, in the COMTEL Conference we found one, and in the ADHO Conference, we found one as well. We accepted 18 articles in total, considering the following criteria that allowed each to be selected.

Inclusion Criteria

- Articles whose methodology has developed and implemented machine translation of languages with low digital resources or distant languages since this systematic review focuses on analyzing the translation methods used for these languages.
- Articles whose title, abstract, and keywords contain the terms: machine translation, low-income languages, and dialect or lingo.
- Articles with publication data from 2016 to 2022.
- Works that are open access and that belong to indexed, refereed journals and conference publications.
- Articles written in English and Spanish.
- Articles from the areas of engineering, computer science, artificial intelligence, and language translation.
- Research articles, short communications, conferences or congresses, magazines, and book chapters.
- An advanced search was performed in the ScienceDirect database in the title, abstract, or author fields with the query string to obtain relevant results.
- In the ScienceDirect database, the filter <<types of articles>> was selected for the following: Review articles, research articles, and book chapters.
- In the ScienceDirect database in areas of knowledge, the following were selected: Medicine and Dentistry, Computer Science, and Engineering.
- Transactions journal articles on Asian and Low-Income Language Information Processing (TALLIP) were selected from the ACM Library database.
- In the ACM database, we used the filter <<Journals>> in the publications section.

Exclusion Criteria

- Articles that do not contain the keywords in the title, abstract, or whose methods are not related to the machine translation of low-resource languages.
- Works of automatic translation of high-resource languages, for example, English-Spanish, French-Spanish, and Spanish-Portuguese.

– After applying the inclusion and exclusion criteria of the articles, we analyzed them in-depth and categorized the results following the phase described below.

2.3 Phase 3: Analysis and Categorization

We performed a quantitative and qualitative analysis of the collected articles in this phase. The first allowed us to analyze the information based on numerical data, while the second allowed us to identify the common characteristics shared by the research works. Each one is detailed below:

Quantitative Analysis
In this quantitative analysis, we took into account the total number of publications that have been made in a period from 2016 to 2022, the countries that have carried out translation work, the areas of knowledge in which these publications are made, what languages of low digital resources have been published and finally the metrics that are implemented in the translation to evaluate the results of the investigations.

Next, Table 1 shows the research questions that we took into account for the quantitative analysis of the collected articles are presented.

Table 1. Questions for the quantitative analysis and motivation.

Questions	Motivations
What is the number of publications that have been made from 2016 to 2022?	Identify publications relevant pear year
What are the countries that publish on the machine translation of low-resource languages?	To identify countries that publish on the machine translation of low-resource language
What areas of knowledge have published articles on the machine translation of languages with low digital resources?	To analyze and identify areas of knowledge that have published articles on the machine translation
In which languages with low digital resources have machine translation been developed?	To identify and show languages with low digital resources where machine translation has been developed
What metrics are implemented in the evaluation of translations from languages with low digital resources?	To identify and analyze metrics implemented in the evaluation of translations from languages with low digital resources
What score did they obtain in the evaluation of the investigations?	To analyze score obtained in the evaluation of the investigation

Results From the Quantitative Analysis
According to the quantitative analysis, the number of publications per year is as follows: in 2016, a total of three documents were obtained; in 2017, no results related to machine

translation were produced; in 2018 and 2019, one article. In 2020 five works were developed; in 2021, seven documents were obtained, and finally, to date, in 2022, there is one article.

Regarding the publications by country, the countries with publications on machine translation of languages with low digital resources are: Germany at 5%, Burma 5%, Canada 5%, China 11%, United States 11%, Finland 5%, Japan 17%, Malaysia 6%, Mexico 17%, Pakistan 6%, South Africa 6%, Tunisia 6%.

According to the systematic review of the articles and the observation of the topics in which they are published in the databases and conferences, we took the following areas of knowledge into account with their respective number of published articles: Medicine and dentistry with zero works, Engineering zero, Computer Science 11, Computational Linguistics six, Social Sciences and Humanities with one. The data indicates that most publications are made in Computer Science and Computational Linguistics. At the same time, no articles related to the translation of languages with low digital resources are reported in the area of Medicine.

Qualitative Analysis

In the qualitative analysis, we took into account the techniques implemented for the translation, the corpus used in the translation works, the scenarios that arise when translating with languages with low digital resources, and the significant results that have been obtained in the investigations, and finally the problems they face when translating these languages. The research questions taken into account for the qualitative analysis of the set of articles collected are shown in Table 2.

Based on the research questions for the qualitative analysis previously mentioned, the following categorization of the works related to machine translation was carried out.

– Translation based on libraries that apply machine learning techniques

Table 2. Questions for qualitative analysis.

Questions	Motivations
What techniques are used in machine translation of low-resource languages?	Identify and categorize techniques used in machine translation of low-resource languages
What kind of corpus are used in a low digital resource environment?	Identify the type of corpus used in a low digital resource environment
What translation scenarios have been used for languages with low digital resources?	Analyze and identify the translation scenarios with existing parallel, monolingual, and not existing corpus
What tools or software are used in the automatic translation of languages with low digital resources?	To categorize tools or libraries used in the automatic translation of languages with low digital resources
What challenges are faced in a translation environment with languages with low digital resources?	To identify challenges in a translation environment with languages with low digital resources

– Translation based on libraries that use statistics.
– Translation based on libraries that make use of grammar rules
– Combination of translation libraries
– Taking into account the classification of the articles below, a brief description of each category is made.

1. Translation Based on Libraries that Apply Machine Learning Techniques
This section describes the works developed using machine learning techniques. Table 3 shows the works whose contents have implemented libraries using machine learning techniques for translation.

In the article by Zacarías and Meza [6], the JoeyNMT[1] library is implemented for automatic translation between the Ayuuk and Spanish languages[2]. For its development, the following steps were followed: automatic alignment, tokenization, orthographic normalization of the corpus, and training with the JoeyNMT tool. A BLEU (Bilingual Evaluation Understudy) above 5.0 was obtained. BLEU is a method of automatic evaluation of machine translation, quickly and language-independent. BLEU measures the quality of translation with respect to a reference [24].

In the work of Knowles et al. [7] highlight the translation of Spanish, Wixárika, Nahuatl, Rarámuri, and Guarani. The Sockeye[3] library was implemented in the translation experiments. The results were validated with the ChrF metric (character n-gram F-score) obtained in Guaraní (gn) 0.258, Wixarika (hch) 0.262, Nahuatl (nah) 0.252, Raramuri (tar) 0.134. ChrF is a technique for the measure of machine translation with the use of the character n-gram F-score. ChrF is language-independent, and tokenization-independent of language [25].

In the work of Mager et al. [8] propose shared tasks for the translation of parallel corpora in the languages Quechua-Spanish, Wixarika-Spanish, Shipibo-Konibo-Spanish, Asháninka-Spanish, Rarámuri-Spanish, Nahuatl-Spanish, Otomí-Spanish, Aymara-Spanish, Guarani-Spanish, and Bribri-Spanish. Researchers can choose to use the baseline that was developed with the FairSeq[4] tool or implement whatever techniques they deem appropriate. The Helsinki method is used, and the results are validated with the ChrF metric of the languages Aymara (aym) 28.3, Bribri(bzd) 16.5, Ashaninka (cni) 25.8, 33.4, Wizarika (hch) 30.04, Nahuatl (nah) 26.6, Otomí 14.7, Quechua (quy) 34.6, Shipibo-Konibo (shp) 32.9, Raramuri (tar) 18.4.

In the article by Vazquez et al. [9], they report a machine translation system based on the OpenNMT[5] tool in combination with pre-training and back-translation. The languages taken into account are Ashaninka, Wixarika and Shipibo-Konibo. The results are validated with the ChrF2 metric obtaining for: Ashaninka(cni) 0.258, Aymara(aym) 0.283, Bribri (bzd) 0.165, Guaraní (gn) 0.336, Hnahñu (oto) 0.147, Nahuatl (nah) 0.266, Quechua (quy) 0.343, Rarámuri (tar) 0.184, Shipibo-Konibo (shp) 0.329, Wixarika (hch) 0.304.

[1] https://github.com/joeynmt/joeynmt
[2] https://github.com/DelfinoAyuuk/corpora_ayuuk-spanish_nmt
[3] https://github.com/awslabs/sockeye
[4] https://github.com/facebookresearch/fairseq
[5] https://opennmt.net

The work presented by Zheng et al. [10] highlights the implementation of the FairSeq library for translation between the languages Aymara, Bribri, Asháninka, Guarani, Wixarika, Nahuatl, Hñähñu, Quechua, Shipibo-Konibo, Rarámuri, Bulgarian, English, French, Irish, Korean, Latin, Spanish, Sundanese, Vietnamese, and Yoruba. Pre-training was performed with mBart-multilingual encoder-decoder (sequence-to-sequence), Tokenization with SentencePiece, and training with FairSeq. Regarding the results, a BLEU of 1.64 and a ChF of 0.0749 were obtained.

The authors Ahmadnia et al. [11] propose machine translation with neural networks in combination with alignment and filtering for Persian-Spanish languages. The bilingual texts used during the training process were taken from the Tanzil corpus, which contains 67 thousand pairs of Persian-Spanish sentences. The use of these filtering techniques considerably improves the results of the translation process. A BLEU of 26.02 was obtained.

Ghafoor et al. [12] report translation using the Google Translate API tool in combination with error analysis. Translations were done between the languages English, Urdu, German, and Hindi. Regarding the results, it was obtained that the accuracy of the English language using SVM (Support vector machine) is 90.45%, and the German data set is 90.01%. In the Urdu language with SVM, an accuracy of 87.26% was obtained, while in the Hindi language with the use of a Bi-LSTM, an accuracy of 85.99% was achieved.

Nekoto et al. [13] publicize machine translation with a community of researchers who share growth strategies, knowledge exchange, and the development of translation models in more than 30 African languages. The JoeyNMT library was implemented in the translation experiments. The evaluation of the translation tool with the support of human experts is highlighted. As future works stand out, to continue with the compilation of parallel corpora, to carry out developments for other areas of natural language processing, and support the tool's implementation for other languages. The following results were obtained using the BLEU metric: Dendi (ddn) 22.30, Pigdin (pcm) 23.29, Fon (fon) 31.07, Luo (luo) 34.33, Hausa (ha) 41.11, Igbo (ig) 34.85,Yoruba (yo) 38.62, Shona (sn) 30.84, Kiswahili (sw) 48.94.

The authors Imankulova et al. [14] translate between Russian, Japanese, French and Malagasy, German and English languages. The OpenNMT library was used for the translation experiments. They propose unsupervised translation to generate bilingual resources to be reused in supervised tasks. They emphasize that the accuracy of the results depends on the length and quality of the training. In future works, the proposal can be trained with the same domain and corpus size, and reinforcement learning can be incorporated. It was obtained with the French-Malagasy with the metric AAS (Average Alignment Similarity) 16.87, Japanese-Russian a BLEU 13.20 was obtained. German-English BLEU 24.13.

2. Translation Based on Libraries that Use Statistics

This section describes the articles whose contents implemented translation libraries with statistics. Table 4 shows the most significant works that were taken into account for this category.

Table 3. Translation with machine learning libraries.

Authors and references	Problem-solving technique	Libraries used	Challenges in language translation	Metrics
Zacarías y Meza [6]	Neural networks	JoeyNMT	Orthographic normalization, Shortage of corpus	BLEU
Knowles et al. [7]	Neural networks	Sockeye	Dialectal and orthographic variety	ChrF
Mager et. al. [8]	Neural networks	FairSeq	Shortage of corpus, Orthographic rules and normalization, Dialectal variety	ChrF
Vazquez et al. [9]	Pre-training, Back translation, Neural networks	OpenNMT	Shortage of corpus, Orthographic normalization	ChrF2
Zheng et al. [10]	Pre-training, Neural networks,	FairSeq	Scarcity of resources, Morphological complexity	ChrF
Ahmadnia et al. [11]	Neural networks	Library not specified	Scarcity of corpus, Morphological complexity	BLEU
Ghafoor et al. [12]	Neural networks	Google Translate API	Scarcity of corpus	Precision
Nekoto et al. [13]	Neural networks, evaluation by human expert	JoeyNMT	Shortage of corpus, language standardization, Difficult adaptation of existing methods, Infrastructure and time limitations	BLEU
Imankulova et al. [14]	Neural networks	OpenNMT	Scarcity of corpus	AAS, BLEU

In the work of Mager et al. [15], Translation between the Wixarika languages and Spanish was developed. The use of the Moses[6] library with the probability technique in combination with morphological segmentation is highlighted. It faces the problem of morphological complexity of the Wixarika language. The result was obtained with WER (Word Error Rate) 38, TER (Translation Error Rate) 0.84. TER measures the amount of

[6] http://www2.statmt.org/moses/

correcting a human expert would have to modify the output to match a reference [25]. On the other hand, WER [26] reduces the word error rate in textual summaries of spoken languages.

Pa et al. [16] highlight the implementation of the Moses tool in translation. Translation comparisons were performed with probability techniques and hierarchical phrase strategies. The latter helps in reordering the words during translation. In addition to the language syntax technique that is built into Moses. The parallel corpus is written in the languages Lao, Myanmar, and Thai. It was obtained with BLEU metric Myanmar-English 21.65, Thai-English 36.98, Lao-English 31.47.

Table 4. Translation with statistics libraries.

Authors and references	Problem-solving technique	Libraries used	Challenges in language translation	Metrics
Mager et al. [15]	Probability technique	Moses	Morphological complexity, Orthographic normalization, Scarcity of corpus	WER, TER
Pa et al. [16]	Probability technique	Moses	No challenges specified	BLEU

3. Translation with Libraries Based on Grammar Rules

The article by [17] discloses the translation with grammatical rules for Arabic languages. The Apertium tool is implemented in the translation since the Apertium[7] library already has the language incorporated in its translation. Future works include continuing to add data to the bilingual dictionary, combining the method with statistics, and adding more semantic rules. In the evaluation, it was obtained with the metrics WER 23.28%, TER 23.85%, and BLEU 55.22.

Table 5. Translation based on grammar rules.

Authors and references	Problem-solving technique	Libraries used	Challenges in language translation	Metrics
Sghaier y r Zrigui [17]	Grammatical rules	Apertium	Morphological and lexical disambiguation	WER, TER, BLEU

[7] https://wiki.apertium.org/wiki/Main_Page

4. Combination of Translation Libraries

Maimaiti et al. [18] proposed the implementation of the THUMT[8] neural network library in combination with transfer learning and word embedding (Word Embeddings of the English language). Applying this combination to low-resource languages helps to find better performance. This method can be applied to other areas of Natural Language Processing (NLP) and other languages as it is language and architecture-independent. It was obtained with BLEU with Azerbaijani and Uzbek languages 4:94 and 4:84.

In the work of [19], they disclose the machine translation with the unsupervised neural network with the Marian[9]library and Moses statistical translation. The corpus that they generated with unsupervised neural translation was reused for supervised tasks. Using supervised and unsupervised neural networks considerably improves the translation quality compared to previous works. In this work, tests were carried out in 5 languages: English, French, German, Indonesian, and Japanese. Regarding the evaluation for Japanese-English languages, a BLEU of 3.9 was obtained, while the Japanese-Indonesian of 0.3.

Yeong et al. [20] propose the use of the Moses library in English-Malay translation experiments in combination with an English language stemmer to improve the translation. Future work highlights the implementation of the Giza++ tool for automatic alignment at the word level. It was obtained with Malay-English BLEU 12.90.

In the article by [21], they expose a scenario with low-resource languages in which there is not a large amount of parallel data and easy access. Neural machine translation training with the OpenNMT library and statistical machine translation with the Moses library in combination with the multidomain corpus were performed. In the evaluation with the BLEU metric with the Gnome corpus, the following results were obtained: Moses 20.54, OpenNMT 15.49, Moses adapted 17.26, OpenNMT adapted 18.76. With the Subtitles corpus, the following scores were obtained: Moses 18.82, OpenNMT 18.62, Moses adapted 19.51, OpenNMT adapted 22.54, respectively.

In the work of [22], they perform the translation between the Chinese and Vietnamese languages. They propose the use of back-translation in combination with the OpenNMT neural network library and Moses statistical translation. Future works include adding more data to the corpus and incorporating transfer learning. According to the results with a focus on Chinese-Vietnamese characters with the METEOR metric with statistical techniques 30.29 and neural networks 25.32.

In the work of Mager and Meza [23], they highlight a comparison between scenarios, on the one hand, those that use the Moses library based on the probability technique. On the other hand, there is the scenario with neural networks with the OpenNMT library. These two libraries were combined with automatic word alignment with the Giza++ library. It faced the challenges of scarce digital resources; languages have morphological complexity, and they do not have orthographic normalization. In this work, experiments were carried out with 5 indigenous languages, Wixarika, Nahuatl, Yorem Nokki, Purépecha, and Mexicanero. Regarding the results, the BLEU metric was obtained with neural networks in the following languages: Mexicanero-Spanish 2.95,

[8] https://github.com/THUNLP-MT/THUMT

[9] https://marian-nmt.github.io

Nahuatl-Spanish 3.04, Purépecha-Spanish 0, Wixarika-Spanish 0, Yorem Nokki-Spanish 0. On the other hand, with the probability technique with the languages: Mexicanero-Spanish 23.47, Nahuatl-Spanish 10.14, Purépecha-Spanish 5.38, Wixarika-Spanish 0, Yorem Nokki-Spanish 2.44.

Table 6. Combination of translation libraries.

Authors and references	Problem-solving technique	Libraries used	Challenges in language translation	Metrics
Maimaiti et al. [18]	Neural networks	THUMT		BLEU
Marie y Fujita [19]	Neural networks, probability techniques	Moses, Marian	Corpus shortage	BLEU
Yeong et al. [20]	Probability techniques, English language lemmatizer	Moses	Corpus shortage	BLEU
Ahmadnia y Dorr [21]	Neural networks, Probability techniques	OpenNMT, Moses	Corpus shortage	BLEU
Li, Sha y Shi [22]	Neural networks and probability techniques	OpenNMT, Moses	Corpus shortage	METEOR
Mager y Meza [23]	Probability techniques and neural networks	OpenNMT, Moses	Scarcity of resources, Morphological complexity, Orthographic normalization	BLEU

2.4 Phase 4: Discussion

According to the results of the analysis of the articles and the characteristics that they share in common, these works were categorized into the following categories:

- Translation with machine learning libraries
- Translation with statistics libraries
- Translation with libraries based on grammar rules
- Combination of translation libraries

According to the analysis of the articles, the following challenges or limitations were generally identified:

- The scarcity of the corpus. It refers to the fact that there is not a considerable amount of digital corpus to support automatic translation tasks.
- Orthographic normalization or standardization in languages. It refers to the fact that languages with low digital resources do not have documentation of the common writing standards for all the variants of the language, which causes inconveniences when translating these languages.
- The morphological complexity of languages. It refers to the fact that languages with low digital resources have many morphemes that can cause difficulties when dealing with these languages.
- The complexity of existing techniques. It refers to the fact that the libraries and tools that already exist for machine translation to apply to languages with low digital resources require an adaptation of the existing tools to languages with low digital resources.
- Computational infrastructure and time. It is mentioned that sometimes there is not enough computer equipment that supports translation tasks, mainly computers, memories, GPUs, and hard drives. In addition to the time limitation, sometimes there is not enough time for machine translation tasks with languages with low digital resources.

Table 3 shows the articles related to translation with machine learning libraries. In the work of Zacarías and Meza [6] in the evaluation, a BLEU above 5.0 was obtained. The translation focuses only on the languages Spanish and Ayuuk. This work used a corpus size of more than 6000 phrases. In this article, we only worked with the corpus of the San Juan Güichicov variant. The authors point out that it is easier to translate from Spanish to Ayuuk than vice versa. The work highlights that the use of Transformers neural networks in low-resource languages yields results, although they point out that the results are low compared to machine translation standards. In the future works of this article, it is highlighted to continue with the compilation of the corpus of other Ayuuk variants, to incorporate the morphological analysis to support the translation, and to continue with the orthographic normalization. However, a high score was obtained in the article by Ahmadnia et al. [11] with a BLEU value of 26.02; this indicates that the quality of the translation and the scores are close to the reference translation, in addition to the fact that in this last article the size of the training corpus is 67K sentence pairs are greater than in the work of Zacarías and Table [6]. On the other hand, the tool implemented in the translation is not mentioned since its own development was carried out with neural network techniques. In this work, he focuses on the Persian and Spanish languages.

In the work of Mager et al. [8] highlight the implementation of the FairSeq library in the translation of low-resource languages in America. Regarding the results with the ChrF metric, the following scores were obtained: Aymara (aym) 0.157, Bribri (bzd) 0.68, Ashaninka (cni) 0.102, Guaraní (gn) 0.193, Wixárika (hch) 0.126, Nahuatl (nah) 0.157, Otomí 0.054, Quechua (quy) 0.304, Shipibo-Konibo (shp) 0.121, Raramuri (tar) 0.039.

The corpus[10] size used is 228275 sentence for training, 9122 sentence for validation and 10018 sentence for test. Although the work presented by Zheng et al. [10] highlights better results with the ChrF metric in the Aymara 0.209, Bribri 0.131, Asháninka 0.214, Guaraní 0.254, Wixarika 0.229, Nahuatl 0.238, Hñähñu 0.133, Quechua 0.33, Shipibo-Konibo 0.175 and Rarámuri 0.123 languages, respectively. In this article, the FairSeq library was implemented for translation and the corpus size used is 13 GB monolingual phrases data, 140 MB phrases of parallel corpus. The authors attribute the improvement of the score to the use of pre-training of the language model, which allowed learning of the languages involved before adjustments. In future works, they propose pre-training using the dictionary augmentation technique, pseudo-monolingual data, and experiments with a probabilistic morphological segmenter of finite states. On the other hand, with the ChrF metric, outstanding scores are obtained in the article by Vázquez et al. [9] in the languages Aymara 0.283, Bribri 0.165, Ashaninka 0.258, Guarani 0.336, Wixárika 0.304, Nahuatl 0.266, Otomí 0.147, Quechua 0.343, Shipibo-Konibo 0.329, Raramuri 0.184. The results improve with the implementation of the OpenNMT translation library, pre-training, and corpus data filtering. In this work, the corpus size is 228274 sentences for training and 9122 sentences for development.

In the article, Knowles et al. [7] highlight the translation with the Sockeye library in languages with low digital resources, in this case, Guaraní, Wixarika, Nahuatl, and Rarámuri. Therefore, this article only focuses on these four languages. Regarding the results with the ChrF metric, the following scores were obtained: Guaraní 0.258, Wixarika 0.262, Nahuatl 0.252, and Rarámuri 0.134. The improvement of scores is tokenization, pre-training, and translation memories. The size corpus used is 65863 sentence pairs for training and 3656 sentence pairs for development.

In the work of Nekoto et al. [13] with the COVID corpus in the HBLEU metric, the following scores were obtained: Dendi (ddn) 0.27, Pidgin (pcm) 3.03, Fon (fon) 15.43, Luo (luo) 0, Hausa (ha) 26.96, Igbo (ig) 11.94, Yoruba (yo) 85.92, Shona (sn) 31.31, Kiswahili (sw) 0 while the following scores were obtained with the TED corpus with the HBLEU metric: Dendi (ddn) 0, Pidgin (pcm) 9.76, Fon (fon) 0, Luo (luo) 7.90, Hausa (ha) 20.42, Igbo (ig) 33.74, Yoruba (yo) 49.22, Shona (sn) 0, Kiswahili (sw) 60.47. These data indicate that in the COVID corpus in the Fon and Igbo languages, the scores are close to each other, while the TED corpus in the Dendi, Fon and Shona languages do not mention the scores obtained. In this work, he focuses on African languages. The corpus size used is 2528078 sentences for training[11].

Table 4 shows the automatic translation works with statistical libraries. Mager et al. [15] present the first translation tool in the Wixarika language with the Moses library. In this work, he focuses on the Wixarika and Spanish languages. Future works include improving the translator with bilingual lexical extraction and continuing with the compilation of the corpus. Regarding the evaluation, it was obtained with WER metrics (without morphological segmentation 38, with morphological segmentation 25, segmentation with labeling 21) and with TER metrics (segmentation 0.84, with morphological segmentation 0.46, segmentation with labeling 0.46) these results indicate that the use

[10] https://github.com/AmericasNLP/americasnlp2021/tree/main/data
[11] https://github.com/masakhane-io/masakhane-mt

of morphological segmentation and labeling considerably improve the results since the error score is lower when implementing these techniques.

On the other hand, the work of Pa et al. [16] explore statistical machine translation in Myanmar, Thai, Lao, and English languages. In the evaluation with the BLEU metric with a statistical technique based on phrases, the following scores were obtained: English-Lao 20.87, Lao-English 31.41, English-Myanmar 10.71, Myanmar-English 21.65, English-Thai 37.33, Thai-English 36.98 while using the hierarchical phrase-based technique, English-Lao 18.94, Lao-English 30.73, English-Myanmar 12.53, Myanmar-English 20.95, English-Thai 38.60, Thai-English 35.45 were obtained. According to these scores, the Thao-English and English-Thao languages received outstanding scores. However, with the translation technique based on hierarchical phrases, the minimum score was 12.53 in the English-Myanmar languages. The corpus size used in Myanmar is 13042 sentence pairs and Lao 35125 words.

Table 5 shows the machine translation jobs with grammar rules. In the article by Sghaier and Zrigui [17], they report the results of the translation with the Apertium library in the Tunisian dialect to Arabic. The following scores were obtained in the evaluation: WER 23.28%, TER 23.85%, and BLEU 55.22. According to the authors, the scores indicate that the error rate was acceptable since it is below 30%, and the BLEU metric indicates that the performance was good since it was a percentage above 50. In this work, the corpus size is 763 words for test and 805 words for reference corpus.

Table 6 shows the jobs that perform the combination of libraries. In the article by Ahmadnia and Dorr [21], they report the results of machine translation in a multi-domain scenario for languages with low digital resources. It focuses on the performance of the Moses library and OpenNMT. In the evaluation with the BLEU metric with the Gnome corpus with Moses 20.54, OpenNMT 15.49, Moses adapted 17.26, OpenNMT adapted 18.76, while with the Subtitles corpus with Moses 18.82, OpenNMT 18.62, Moses adapted 19.51, OpenNMT adapted 22.54. These data indicate that better results are obtained with Moses than with OpenNMT. Although Moses and OpenNMT adapted to a specific domain, the latter's results improved. The corpus sized was 5213125 sentences.

In the work by Mager and Meza [23], they present the advances in machine translation for five low-resource languages. In this work, they compare the Moses and OpenNMT libraries. They emphasize that using morphological segmentation improves the results with both libraries. In the evaluation with the BLEU metric with the Moses library, the following values were obtained: Mexicanero-Spanish 23.47, Nahuatl-Spanish 10.14, Purépecha-Spanish 5.38, Wixárika-Spanish 2.44, Yorem Nokki-Spanish 0, while with the OpenNMT library the following values were obtained: obtained the following scores: Mexicanero-Spanish 2.95, Nahuatl-Spanish 3.04, Purépecha-Spanish 0, Wixárika-Spanish 0, Yorem Nokki-Spanish 0. The data indicate that the results of the Moses library are better than the OpenNMT library because they are trained with a small corpus. Furthermore, the Mexicanero and Nahuatl languages performed better than the Wixarika language, considering that the Wixarika language has a greater number of morphemes per word than Nahuatl. The corpus size used is 985 sentence pairs.

3 Conclusions

According to the searches carried out in databases such as ScienceDirect, ACM Library, and IEEE, there are few articles on machine translation of low-resource languages in Mexico. According to the review, most of them have been published at international conferences, in addition to the fact that, in Mexico, there are few works on rule-based machine translation with languages with low digital resources due to the complexity of creating language rules. Also, through the systematic review, we identified that in the area of medicine, articles on the automatic translation of languages with low digital resources were not published, making it difficult to accurately communicate information on the health of patients from indigenous communities. So this area may be an area of opportunity in future work. In Mexico's machine translation advances were identified for the following indigenous languages: Nahuatl, Wixarika, Otomí, Raramuri, Ayuuk, Yorem Nokki, Purépecha, and Mexicanero. According to our research in databases and conferences, there are no automatic translation works on Mixteco, Amuzgo, Tlapaneco, Chatino, among others.

References

1. Gutierrez-Vasques, X., Vilchis-Vargas, E., Cerbon-Ynclan, R.: Recopilación de un corpus paralelo electrónico para una lengua minoritaria: el caso del español-náhuatl. In: Primer Congreso Internacional el Patrimonio Cultural y las Nuevas Tecnologías. Ina. (2015)
2. Hedderich, M.A., Lange, L., Adel, H., Strötgen, J., Klakow, D.: A survey on recent approaches for natural language processing in low-resource scenarios. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2545–2568 (2021)
3. Palacios-Díaz, R., Escudero-Nahón, A.: Revisión sistemática de los desafíos del uso de tecnología digital en la formación de investigadores. Educateconciencia **26**(27), 147–178 (2020)
4. Singh, A.K.: Natural language processing for less privileged languages: where do we come from? Where are we going? In: Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pp. 7–12, January 2008
5. Cieri, C., Maxwell, M., Strassel, S., Tracey, J.: Selection criteria for low resource language programs. In: Proceedings of the 10th International Conference on Language Resources and Evaluation, Lr. 2016, pp. 4543–4549 (2016)
6. Zacarías Márquez, D., Meza Ruiz, I.V.: Ayuuk-Spanish neural machine translator, pp. 168–172 (2021)
7. Knowles, R., Stewart, D., Larkin, S., Littell, P.: NRC-CNRC machine translation systems for the 2021 AmericasNLP shared task. In: Proceedings of the 1st Workshop on Natural Language Processing for Indigenous Languages of the Americas 2021, pp. 224–233 (2021)
8. Mager, M., et al.: Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas, pp. 202–217 (2021)
9. Vázquez, R., Scherrer, Y., Virpioja, S., Tiedemann, J.: The Helsinki submission to the AmericasNLP shared task. In: Proceedings of the 1st Workshop on Natural Language Processing for Indigenous Languages of the Americas 2021, pp. 255–264 (2021)
10. Zheng, F., Reid, M., Marrese-Taylor, E., Matsuo, Y.: Low-resource machine translation using cross-lingual language model pretraining, pp. 234–240 (2021)

11. Ahmadnia, B., Dorr, B.J., Aranovich, R.: Impact of filtering generated pseudo bilingual texts in low-resource neural machine translation enhancement: the case of Persian-Spanish. Procedia CIRP **189**, 136–141 (2021)
12. Ghafoor, A., et al.: The impact of translating resource-rich datasets to low-resource languages through multi-lingual text processing. IEEE Access **9**(2), 124478–124490 (2021)
13. Nekoto, W., et al.: Participatory research for low-resourced machine translation: a case study in African languages, pp. 2144–2160 (2020)
14. Imankulova, A., Sato, T., Komachi, M.: Filtered pseudo-parallel corpus improves low-resource neural machine translation. ACM Trans. Asian Low Resour. Lang. Inf. Process. **19**(2), 1–16 (2019)
15. Mager Hois, J.M., Barrón Romero, C., Meza Ruiz, I.V.: Traductor estadístico wixarika-español usando descomposición morfológica. COMTEL, pp. 63–68 (2016)
16. Pa, W.P., Thu, Y.K., Finch, A., Sumita, E.: A Study of statistical machine translation methods for under resourced languages. Procedia Comput. Sci. **81**, 250–257 (2016)
17. Sghaier, M.A., Zrigui, M.: Rule-based machine translation from Tunisian dialect to modern standard Arabic. Procedia Comput. Sci. **176**, 310–319 (2020)
18. Maimaiti, M., Liu, Y., Luan, H., Sun, M.: Enriching the transfer learning with pre-trained lexicon embedding for low-resource neural machine translation. Tsinghua Sci. Technol. **27**(1), 150–163 (2022)
19. Marie, B., Fujita, A.: Iterative training of unsupervised neural and statistical machine translation systems. ACM Trans. Asian Low Resour. Lang. Inf. Process. **19**(5), 1–21 (2020)
20. Yeong, Y.L., Tan, T.P., Mohammad, S.K.: Using dictionary and lemmatizer to improve low resource English-Malay statistical machine translation system. Procedia Comput. Sci. **81**, 243–249 (2016)
21. Ahmadnia, B., Dorr, B.J.: Low-resource multi-domain machine translation for Spanish-Farsi: neural or statistical? Procedia Comput. Sci. **177**, 575–580 (2020)
22. Li, H., Sha, J., Shi, C.: revisiting back-translation for low-resource machine translation between Chinese and Vietnamese. IEEE Access **8**, 119931–119939 (2020)
23. Mager, M., Meza, I.: Hacia La Traducción Automática De Las Lenguas Indígenas De México. Digital Humanities 2018, México City, pp. 1–7 (2018)
24. Papineni, K., Roukos, S., Ward, T., . Zhu, W.: BLEU: a method for automatic evaluation of machine translation. In: Proceeding of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)
25. Papovic, M.: chrF: chareacter n-gram F-score for automatic MT evaluation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation, pp. 392–395 (2015)
26. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of the 7th Conference of Association for Machine Translation in the Americas: Technical papers, Cambridge, Massachsetts, USA, pp. 223–231 (2006)
27. Zechner, K., Waibel, A.: Minimizing word error in textual summaries of spoken language. In: Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference, pp. 186–193 (2000)