Improving Low-Resource Machine Translation for Formosan Languages Using Bilingual Lexical Resources

Francis Zheng, Edison Marrese-Taylor, Yutaka Matsuo
Graduate School of Engineering
The University of Tokyo
{francis, emarrese, matsuo}@weblab.t.u-tokyo.ac.jp

Abstract

This paper investigates how machine translation for low-resource languages can be improved by incorporating information from bilingual lexicons during the training process for mainly translation between Mandarin and Formosan languages, which are all moribund or critically endangered, and we also show that our techniques work for translation between Spanish and Nahuatl, a language pair consisting of languages from completely different language families. About 70% of the approximately 7,000 languages of the world have data in the form of lexicons, a valuable resource for improving low-resource language translation. We collect a dataset of parallel data and bilingual lexicons between Mandarin and 16 different Formosan languages and examine mainly three different approaches: (1) simply using lexical data as additional parallel data, (2) generating pseudo-parallel sentence data to use during training by replacing words in the original parallel sentence data using the lexicon, and (3) a combination of (1) and (2). All three approaches give us gains in both BLEU scores and CHRF scores, and we found that (3) provided the most gains, followed by (1) and then (2), which we observed for both translation between Mandarin and the Formosan languages and Spanish-Nahuatl. With technique (3), we saw an average increase of 5.55 in BLEU scores and 10.33 in CHRF scores.

1 Introduction

Evidence shows that neural machine translation (NMT) systems do not work well with languages that lack parallel data (Koehn and Knowles, 2017), and typically, such machine translation systems require millions of parallel sentences as training data, which are only available for a limited number of language pairs (Haddow et al., 2022). However, there are several low-resource languages with a significant number of speakers, such as Burmese and (Haddow et al., 2022), highlighting the need for more attention to be paid to the development of machine translation systems for these languages. For example, while there are nearly 280 million English-French parallel sentences in OPUS (Tiedemann, 2012), there are only about 700,000 English-Burmese parallel sentences, even though Burmese (Myanmar) has tens of millions of speakers (Haddow et al., 2022). Furthermore, it is important to improve machine translation for all languages, even those with few speakers, to help provide fair access to technology.

Though parallel data are only available for a limited number of language pairs, bilingual lexicons are an often overlooked resource that are more readily available and less expensive to produce (Haddow et al., 2022). Language documentation efforts typically produce bilingual lexicons first, and about 70% of the approximately 7,000 languages of the world have such lexicons (Wang et al., 2022). They may be advantageous in that they can provide information about infrequent terms that may not appear in parallel data, but they also are not as helpful as parallel data in that they do not give information about how a word may be translated differently depending on the context (Haddow et al., 2022). However, despite the lack of contextual information, bilingual lexicons still provide additional valuable information that machine translation systems can use.

This paper explores how bilingual lexicons can be used as a complement to parallel data in the context of low-resource machine translation for Asian languages. In contrast to prior work which so far has mostly used replacement-based techniques which leverage bilingual lexicons by generating pseudo-parallel data, in this paper we propose to simply use lexical data as additional parallel data for training. To test our idea, we introduce a dataset containing parallel sentences and bilingual lexicons for 16 Formosan languages, which have a great deal of diversity in linguistic struc-

Findings of the Association for Computational Linguistics: ACL 2024, pages 11248–11259
August 11-16, 2024 ©2024 Association for Computational Linguistics

ture, which we use to propose an extensive empirical framework. Concretely, we aim to answer the following research questions: (RQ1) Can adding lexical information as parallel data improve translation quality?, (RQ2) How much lexical information is needed to improve translation quality?, (RQ3) How does using lexical information compare with using parallel data?, (RQ4) How does using lexical information compare with pseudo-parallel data?

Our results demonstrate that adding lexical information during the training process can produce significant improvements in translation quality for real-world, low-resource language pairs, using Formosan languages and Mandarin as an example. We also examine the relationship between improvements in translation quality and the amount of additional lexical information provided, and compare the effect of using lexical information with the effect of using additional parallel sentence data. Additionally, we also demonstrate that our methods improve translation quality for one non-Asian language pair, Spanish-Nahuatl, suggesting that our technique may generalize to other language pairs.

2 Related Work

Bilingual lexicons have proven useful in aiding a large a variety of downstream tasks such as semantic role labeling (van der Plas et al., 2011), sentence alignment (Fernando et al., 2023), and part-of-speech tagging (Wang et al., 2022; Täckström et al., 2013). Wang et al. (2022) explored the use of bilingual lexicons in three tasks: named entity recognition, part-of-speech tagging, and dependency parsing. The authors used bilingual lexicons to synthesize data and were able to use these synthesized data to achieve significant improvements in performance in all three of these tasks across 19 under-represented languages. These synthesized data were made by using bilingual lexicons to create synthetic sentences in the target language through word-to-word translation. Word order and morphological differences between the source and target languages were not considered in this process, so the synthetic data may be quite different from true data in that language (Wang et al., 2022).

Bilingual lexicons have also been used to improve machine translation. Despite the fact that approximately 70% of the world's languages have documentation in the form of bilingual lexicons (Wang et al., 2022), techniques involving bilingual lexicons in machine translation typically involve artificially generated lexicons.

Dinu et al. (2019) used terms extracted from two English-German lexicons and inserted these terms after their corresponding source words to improve translation quality. Prior to this work, most work on neural machine translation used constrained decoding to integrate terminology, but Dinu et al. (2019) used a *black-box* generic neural machine translation architecture that is directly trained to use an external lexicon provided during run-time, achieving increases of 0.2 to 0.9 in BLEU scores.

A more recent multilingual pre-trained method called mRASP (Lin et al., 2020) also makes use of lexicons to improve translation, as demonstrated by experiments on 42 translation directions, including both low-resource and high-resource languages. In this method, a lexicon composed of unsupervised word alignments generated by MUSE (Conneau et al., 2017) is used to substitute words in the source sentence during pre-training. This is done to bring semantically similar words from different languages closer in the representation space (Haddow et al., 2022). Through a set of analytical experiments, the authors also showed that alignment information does help bridge any gaps between languages and improve translation quality.

AFROMT (Reid et al., 2021) is a machine translation benchmark for eight African languages, and the authors of this benchmark demonstrated that by extracting a dictionary from parallel corpora using `eflomal`[1] (Östling and Tiedemann, 2016) and taking word alignments that appear over 20 times to produce a bilingual lexicon, this lexicon can be used to produce code-switched monolingual data to aid in training machine translation models. The authors also experimented with iteratively creating pseudo-monolingual data in low-resource languages.

Most recently, Jones et al. (2023) showed that using bilingual lexicons to substitute source sentence words for their translations to create code-switched data and prepending translations of source words to source sentences are two techniques that can improve machine translation quality for low-resource and unsupervised languages. The authors used bilingual lexicons from PanLex

[1] https://github.com/robertostling/eflomal/

(Kamholz et al., 2014) to do this.

Our work shows that there may be an even simpler approach to integrating lexical information when training neural machine translation systems. This makes use of existing bilingual lexicons without the need for inferring a bilingual lexicon from parallel data.

3 Data

To explore whether lexicons can be simply used as additional parallel data in a low-resource setting in the context of Asian languages, we believe it is important to experiment with a diverse set of languages. To that end, for this paper we collect data from 16 different Formosan languages, a geographic grouping of Austronesian languages indigenous to Taiwan, which have a great deal of diversity in linguistic structure (Li, 2008) despite all being from a relatively small geographic area.

To obtain this data, we downloaded lexicon entries and parallel sentences from an online dictionary (原住民族語言線上辭典)[2] published by the Indigenous Languages Research and Development Foundation (原住民族語言研究發展基金會), an organization based in Taiwan dedicated to research on Taiwan's indigenous languages. We specifically consider the following languages: Amis, Atayal, Bunun, Kanakanavu, Kavalan, Paiwan, Puyuma, Rukai, Saaroa, Saisiyat, Sakizaya, Seediq, Thao, Truku, Tsou, and Yami. The dictionaries consist of words, phrases, and example sentences relating to daily life from each of these languages with their Mandarin equivalents. We manually verified the data don't contain personal information or offensive content, and to the best of our knowledge, our use of these data is compatible with the original access conditions.

While these data are available in an online dictionary format that can be searched, they were not immediately ready for computational use. One can download parts of the dictionary as PDF or ODT files, but neither format is easy to use due to how the lexicon entries are organized. Therefore, we downloaded PDFs from the website of the online dictionary, converted them into HTML using PDFMiner[3], and extracted data using Beautiful Soup[4]. Each Formosan word/phrase and its Mandarin dictionary entry were extracted along with parallel sentence data included in this dictionary. These data are summarized in Table 1 and Table 13. For each language pair, we shuffled the sentences and took 80% of the parallel sentence data to use as our training set, 10% of the data to use as our development set, and the remaining 10% of the data to use as our test set. Our dataset can be found at `https://github.com/francisdzheng/formosan-mandarin`.

Language	Parallel sentences	Lexicon entries
Amis	5,751	7,800
Atayal	5,751	7,337
Bunun	8,975	7,859
Kanakanavu	6,618	5,214
Kavalan	8,216	8,376
Paiwan	5,158	6,664
Puyuma	6,894	9,198
Rukai	10,399	12,806
Saaroa	4,799	6,652
Saisiyat	6,049	7,242
Sakizaya	5,737	7.593
Seediq	5,459	6,723
Thao	7,440	5,716
Truku	4,597	35,056
Tsou	4,437	6,742
Yami	6,483	8,227

Table 1: Summary of the Formosan language data obtained from Taiwan's Indigenous Languages Research and Development Foundation, where each language is paired with Mandarin

Our studied Formosan languages are all written using the Latin script, unlike Mandarin Chinese. Though there are a small number of Formosan languages, their linguistic features are diverse (Li, 2008). For example, word order, focus systems, auxiliaries, numerals, personal pronouns, compounding, affixation, and phonology all demonstrate great diversity across the Formosan languages (Li, 2008). Most Formosan languages are verb-initial, but some also have an SVO (subject-verb-object) order and use different word orders depending on the situation. Thao and Atayal demonstrate little compounding, while Tsou and Bunun demonstrate rich compounding in their vocabularies. Additionally, while some Formosan languages distinguish between human and nonhuman numerals, some also do not. Some Formosan languages require auxiliaries, while others do not. We hope these examples, though limited, help illustrate how linguistically diverse the Formosan languages are.

One key point is that all of our studied languages are moribund or critically endangered. For

[2] `https://e-dictionary.ilrdf.org.tw/index.htm`

[3] `https://github.com/pdfminer/pdfminer.six`

[4] `https://www.crummy.com/software/BeautifulSoup/`

example, Thao was documented to have only four native speakers according to Ethnologue in 2022. Yet some other languages, such as Amis, had roughly 30,000 speakers as of 2015 (Kuo, 2015). This is still, however, a much smaller number of speakers than the most commonly spoken languages in the world, such as English and Spanish, and to that extent, data for Formosan languages are extremely lacking.

It is also important to note that existing literature on these languages written in English may conflate the Truku language and Seediq language. This is due to the fact that the Truku people (also known as the *Taroko* people) were previously grouped with the Seediq people by the Taiwanese government. Truku, in fact, has been described as a dialect of Seediq (Lee et al., 2011), which means that some literature may use *Seediq* as a general term to refer to languages including Truku. The two languages also share the same ISO 639-3 code, *trv*, which may cause further confusion. In this paper, *Truku* refers to what's known as *Tàilǔgé-yǔ* (太魯閣語) in Mandarin, and *Seediq* refers to what's known as *Sàidékè-yǔ* (賽德克語) in Mandarin.

Additionally, to test how our ideas work with an externally created lexicon, we used a Seediq-Mandarin lexicon of 1,556 entries from a Seediq-Mandarin dictionary[5] by Temi Nawi (曾瑞琳). This lexicon was kindly provided to us by Darryl Sterk.

Finally, to explore how our ideas work with non-Formosan languages, we obtained parallel data and a lexicon for Spanish and Nahuatl, a language native to Mexico. Parallel data was obtained from AmericasNLP[6] and originally comes from Axotol (Gutierrez-Vasques et al., 2016), a Spanish-Nahuatl parallel corpus. This dataset consists of 16,145 sentences in the training set, 672 sentences in the development set, and 1,003 sentences in the test set. A Nahuatl-Spanish lexicon containing 10,888 lexicon entries was obtained from AULEX[7]. This lexicon consists of entries from three dictionaries: *Diccionario español-nahuatl* by Manuel Rodríguez Villegas, *Diccionario náhuatl de la huasteca veracruzana - español* by Marcelino Hernández Beatriz, and *Diccionario náhuatl-español* by Francisco Xavier Clavijero and Sybille de Pury.

4 Experimental Setup

As described in Section 2, bilingual lexicons have played a significant role in some of the latest developments in machine translation. In contrast to existing work, we propose to simply treat bilingual lexicons as additional parallel data to explicitly inform the model of translations of different words.

Preprocessing Data were tokenized using the unigram (Kudo, 2018) implementation of SentencePiece (Kudo and Richardson, 2018). A vocabulary size of 8,000 and a character coverage rate of 0.99 were used to train our SentencePiece model.

Models Our Transformer models use six encoder and decoder layers with four attention heads each, a hidden dimension of 512, and a feed-forward size of 1024, and a learning rate of 0.0005. Our model was optimized using Adam (Kingma and Ba, 2015) with hyperparameters $\beta = (0.9, 0.98)$ and $\epsilon = 10^{-8}$. A dropout rate of 0.3 and a weight decay of 0.0001 were used for regularization. Each model for each language pair and direction was trained for two hours on three Quadro RTX 8000 GPUs.

Evaluation Using the SACREBLEU library[8] (Post, 2018), we evaluated the translations outputted by our models with detokenized BLEU (Papineni et al., 2002; Post, 2018) on the test data from our parallel sentence data. We also used CHRF (Popović, 2015) to measure performance at the character level since most Austronesian languages are agglutinative (Blust, 2013), meaning they make extensive use of affixes and are morphologically rich. For example, some languages such as Thao (Blust, 2003) and Tsou (Tsuchida, 1990) have hundreds of prefixes.

5 Results and Analysis

We begin by creating a baseline by training models using only the parallel sentence data and one in which the model was trained using both the parallel sentence data and bilingual lexicon. To incorporate the bilingual lexicon during the training process, we added the lexicon entries in the source lan-

[5] Dictionary title in original language: 賽德克語辭典 (Published by 國家文化藝術基金會)

[6] https://github.com/AmericasNLP/americasnlp2021/tree/main/data/nahuatl-spanish, https://github.com/AmericasNLP/americasnlp2021/blob/main/test_data/test.nah

[7] https://aulex.org/nahuatl/descarga.php?archivo=nau-es

[8] https://github.com/mjpost/sacrebleu

	to Mandarin						from Mandarin					
	Baseline		+ Lexicon				Baseline		+ Lexicon			
Language	**BLEU**	**CHRF**	**BLEU**	**Δ**	**CHRF**	**Δ**	**BLEU**	**CHRF**	**BLEU**	**Δ**	**CHRF**	**Δ**
Amis	3.56	12.08	11.12	+7.56	32.10	+20.02	11.12	32.10	14.72	+3.60	39.22	+7.11
Atayal	4.86	12.26	5.83	+0.97	14.47	+2.21	13.10	35.07	14.09	+0.99	35.36	+0.29
Bunun	5.44	17.91	6.50	+1.06	22.26	+4.36	16.35	40.99	22.24	+5.89	49.73	+8.74
Kanakanavu	9.54	20.93	13.92	+4.38	28.32	+7.39	17.93	39.98	24.24	+6.31	51.05	+11.08
Kavalan	7.18	24.03	8.26	+1.08	30.10	+6.07	29.71	51.98	34.06	+4.35	58.00	+6.02
Paiwan	3.80	4.86	10.64	+6.84	13.63	+8.77	1.73	20.48	9.38	+7.65	33.68	+13.20
Puyuma	7.86	15.69	12.55	+4.69	21.52	+5.83	18.88	43.67	22.31	+3.43	48.33	+4.66
Rukai	8.44	36.66	9.98	+1.54	39.93	+ 3.27	1.19	12.76	2.96	+1.77	15.86	+3.10
Saaroa	6.03	14.06	8.59	+2.56	16.47	+2.41	7.06	33.56	9.28	+2.22	40.25	+6.70
Saisiyat	3.99	16.07	6.62	+2.63	23.08	+7.01	19.37	45.07	24.28	+4.91	50.97	+5.90
Sakizaya	3.11	14.38	5.76	+2.65	20.47	+6.09	12.79	34.48	16.01	+3.22	39.62	+5.13
Seediq	1.52	13.24	2.74	+1.22	15.69	+2.44	8.78	28.02	10.64	+1.86	30.92	+2.91
Thao	10.50	26.66	14.74	+4.24	32.60	+5.94	26.40	50.45	31.48	+5.08	57.25	+6.80
Truku	1.26	6.87	2.58	+1.32	11.20	+4.33	8.08	23.39	12.26	+4.18	30.24	+6.85
Tsou	2.07	19.50	4.37	+2.30	24.24	+4.74	15.61	36.97	17.84	+2.23	41.53	+4.57
Yami	4.72	18.27	6.05	+1.33	23.02	+4.77	20.32	37.84	24.05	+3.73	43.14	+5.29

Table 2: Translation quality in both directions (to and from Mandarin) before adding lexical data ("Baseline") and after adding lexical data ("+ Lexicon") along with changes in translation quality ("Δ")

	0	1,000	2,000	3,000	4,000	5,000	6,000	7,000	8,000	R^2
Amis	3.56	4.98	6.03	7.34	8.01	8.57	9.74	10.98		0.989
Atayal	4.865	4.94	4.87	5.12	5.23	5.11	5.45	5.73		0.831
Bunun	5.44	5.31	5.78	5.82	6.21	6.18	6.25	6.45		0.906
Paiwan	3.8	6.01	7.33	7.4	8.23	9.7	11.07			0.954
Sakizaya	3.11	3.14	4.43	4.12	4.95	5.13	5.01	5.44		0.866
Kavalan	7.18	7.49	7.23	7.51	7.68	7.61	7.82	8.02	8.14	0.885
Rukai	8.44	8.51	8.78	8.59	8.92	9.1	9.34	9.41	9.08	0.791
Puyuma	7.86	7.9	8.43	8.12	8.79	9.54	10.32	11.47	11.69	0.906
Seediq	1.52	1.68	1.62	1.98	2.24	2.46	2.64			0.95
Thao	10.5	11.43	11.78	12.53	13.23	14.18				0.988
Saaroa	6.03	5.85	5.29	6.29	7.39	7.61	7.94			0.75
Yami	4.72	4.9	5.15	5.13	5.48	5.57	5.77	5.98	5.83	0.948
Truku	1.26	1.43	1.4	1.57	1.58	1.62	1.69	1.84	1.92	0.955
Tsou	2.07	2.56	2.93	3.1	3.74	4.12	4.25			0.981
Kanakanavu	9.54	11.43	12.98	13.02	13.91	13.19				0.743
Saisiyat	3.99	4.24	4.53	5.12	5.24	5.93	6.23	6.5		0.986

Table 3: Number of lexicon entries vs. BLEU scores for translation from Formosan languages to Mandarin

guage to the parallel data in the source language and the lexicon entries in the target language to the parallel data in the target language, essentially treating the lexicon entries as additional parallel data.

Table 2 summarizes our results in terms of BLEU and CHRF scores, clearly allowing us to decide whether lexical information can simply be added to our parallel data to improve translation quality. In the table, the "Baseline" columns contain results achieved using only parallel sentence data as training data, and the "+ Lexicon" columns contain results achieved using both parallel sentences and the bilingual lexicon as training data. Overall, we see an average increase of 3.37 in BLEU scores and 6.06 in CHRF scores after adding lexical data during training. Every language pair demonstrated an increase in translation quality regardless of translation direction (to Mandarin and from Mandarin) and evaluation method (BLEU and CHRF). On average, when translating from Formosan languages to Mandarin, we saw an increase of 2.90 in BLEU scores and 5.98 in CHRF scores. When translating from Mandarin to Formosan languages, we saw an average increase of 3.84 in BLEU scores and 6.15 in CHRF scores. Improvements ranged from 0.97 to 7.65 for BLEU scores and 0.99 to 20.02 for CHRF scores.

Having established the effectiveness of our technique, we present a sensitivity study regarding the amount of parallel and lexical data used to train our model, in an attempt to provide insights to

	0	1,000	2,000	3,000	4,000	5,000	6,000	7,000	8,000	R^2
Amis	11.12	11.43	11.51	12.08	12.73	13.51	13.42	14.13		0.962
Atayal	13.1	13.18	13.17	13.52	13.63	13.79	13.85	13.91		0.945
Bunun	16.35	17.52	17.66	18.13	18.8	20.54	19.95	20.97		0.935
Paiwan	1.73	2.51	3.72	6.42	7.14	7.65	9.05			0.965
Sakizaya	12.79	12.92	13.41	13.62	14.05	14.28	14.67	15.42		0.973
Kavalan	29.71	29.73	30.01	30.98	31.13	31.58	32.28	32.88	33.62	0.97
Rukai	1.19	1.22	1.32	1.23	1.48	1.75	1.94	2.12	2.34	0.916
Puyuma	18.88	19.1	19.43	19.49	20.04	20.58	20.17	21.34	21.55	0.926
Seediq	8.78	9.14	9.52	9.61	10.04	10.43	10.49			0.979
Thao	26.4	27.49	27.98	29.15	30.59	31.42				0.986
Saaroa	7.06	7.42	7.99	8.55	7.79	9.04	9.17			0.811
Yami	20.32	20.37	20.41	20.96	21.57	21.83	22.49	23.22	23.79	0.951
Truku	8.08	8.14	8.42	8.49	8.95	9.04	9.43	9.47	10.43	0.922
Tsou	15.61	15.92	16.34	16.53	16.98	17.24	17.52			0.995
Kanakanavu	17.93	18.13	19.43	21.13	23.28	24.18				0.961
Saisiyat	19.37	19.52	20.45	21.18	21.47	22.01	21.94	23.92		0.925

Table 4: Number of lexicon entries vs. BLEU scores for translation from Mandarin to Formosan languages, where R^2 denotes the Pearson correlation between translation quality and number of lexicon entries used during training.

	to Mandarin		**from Mandarin**	
Lex. entries	**BLEU**	**CHRF**	**BLEU**	**CHRF**
0	2.23	10.93	4.59	14.27
1,000	2.45	11.48	5.72	16.40
2,000	3.91	13.49	8.56	19.40
3,000	4.12	14.01	9.43	21.49
4,000	5.34	16.95	11.9	24.25
5,000	6.01	18.43	12.53	26.50

Table 5: Result of adding additional lexicon entries for translation between Mandarin and Kanakanavu

	to Mandarin		**from Mandarin**	
Parallel sents.	**BLEU**	**CHRF**	**BLEU**	**CHRF**
0	2.23	10.93	4.59	14.27
1,000	3.21	11.97	5.29	16.42
2,000	4.65	15.24	7.89	19.03
3,000	6.43	17.40	12.43	25.49
4,000	8.45	18.81	15.79	28.94

Table 6: Result of adding additional parallel sentences for translation between Mandarin and Kanakanavu

	to Mandarin		**from Mandarin**	
Lex. entries	**BLEU**	**CHRF**	**BLEU**	**CHRF**
0	1.09	5.30	3.05	8.47
1,000	1.93	6.18	3.23	9.00
2,000	2.28	7.99	4.05	11.47
3,000	3.51	9.59	5.62	15.88
4,000	4.12	12.49	5.83	16.10
5,000	5.10	15.50	7.42	18.03

Table 7: Result of adding additional lexicon entries for translation between Mandarin and Tsou.

	to Mandarin		**from Mandarin**	
Parallel sents.	**BLEU**	**CHRF**	**BLEU**	**CHRF**
0	1.09	5.30	3.05	8.47
1,000	2.46	7.74	4.67	12.11
2,000	4.02	11.63	5.92	15.09

Table 8: Result of adding additional parallel sentences for translation between Mandarin and Tsou

help answer RQ2 and RQ3. To this end, we experiment with adding different amounts of the lexicon (in increments of 1,000 words) to the parallel sentence data to see the effect on translation quality. Because we did not have additional parallel data for many of these languages, we explored the third question by simply using 1,000 of the original parallel sentences to create a baseline for translation and set aside another 1,000 parallel sentences to use as test data. Then, we compared the effect of adding parallel sentences (in increments of 1,000) and the effect of adding lexical entries (in increments of 1,000) on translation quality. For this experiment, we used translation between Kanakanavu and Mandarin and translation between Tsou and Mandarin as examples.

Tables 3 and 4 summarize our sensitivity study results, showing how much lexical information is needed to attain performance improvements. We see that across all languages and directions in translation, there is a general trend of increases in translation quality as we add more lexicon entries. This is also indicated by the R^2 values shown in the tables, which indicate the correlation between the translation quality achieved and number of lexicon entries used during training.

Results for our sensitivity study regarding parallel data can be found in Tables 5 and 6 for translation between Mandarin and Kanakanavu and Tables 7 and 8 for translation between Mandarin and Tsou. Overall, we observe that more gains in translation quality were achieved by adding parallel sentences compared to adding lexicon entries. For example, for translation from Kanakanavu to Mandarin, adding 4,000 lexicon entries to the training data increased translation quality by 3.11 BLEU points, whereas adding 4,000 parallel sentences to the training data increased translation quality by 6.22 BLEU points.

5.1 Comparison with prior work

As mentioned earlier, previous work has shown that pseudo-parallel data generation approaches (Imankulova et al., 2017; Wang et al., 2022) are very effective in improving translation results for low resource languages. One method of creating such pseudo data is by using a lexicon to replace words in the existing data (Wang et al., 2022). To show the effectiveness of our models and help answer RQ4, we follow Wang et al. (2022) and proceed to create pseudo-parallel data by using the lexicon to replace words in the parallel data. Words were only replaced in the Formosan language data due to tokenization difficulties in the Mandarin data, and we also experimented with different rates of replacement to figure out the optimal rate of replacement. Models were then trained with both the existing parallel data and pseudo-parallel data together (effectively doubling the amount of parallel sentence data). For this experiment, we used translation between Mandarin and six of our Formosan languages, Amis, Bunun, Kanakanavu, Paiwan, Sakizaya, and Seediq, due to computational budget limitations. Additionally, we combine our techniques and train models using the existing parallel data, our pseudo-parallel data, and the lexicons all together.

Table 9 shows the translation quality achieved using pseudo-parallel data in addition to the original parallel data during the training process (indicated in the "+ PPD" rows). For comparison, the translation quality achieved using only the original parallel data and the translation quality achieved using a combination of the original parallel data and lexicon ("+ Lex.") are also shown.

We experimented with pseudo-parallel data created using different rates of replacement, 10%, 20%, 30%, 40%, 60%, and 80%, and found that a 20% rate of replacement produced the best results in most cases. Table 9 displays the best results we achieved across different models trained with the pseudo-parallel data created using different rates of placement. In all cases, using a combination of pseudo-parallel data and lexicon entries proved to produce higher translation quality than using just the lexicon entries as additional data during training.

	to Mandarin		from Mandarin	
Language	BLEU	CHRF	BLEU	CHRF
Amis	3.56	12.08	11.12	32.10
+ Lex.	11.12	32.10	14.72	39.22
+ PPD	11.03	34.25	13.29	37.416
+ Lex. + PPD	15.25	34.25	17.48	42.439
Bunun	5.44	17.91	16.35	40.99
+ Lex.	6.50	22.26	22.24	49.73
+ PPD	6.98	24.44	22.58	49.15
+ Lex. + PPD	8.09	27.49	25.34	55.20
Kanakanavu	9.54	20.93	17.93	39.98
+ Lex.	13.92	28.32	24.24	51.05
+ PPD	12.93	24.83	22.98	48.01
+ Lex. + PPD	14.92	29.35	25.43	54.29
Paiwan	3.80	4.86	1.73	20.48
+ Lex.	10.64	13.63	9.38	33.68
+ PPD	6.48	11.50	7.98	31.04
+Lex. + PPD	10.94	15.39	9.46	35.49
Sakizaya	3.11	14.38	12.79	34.48
+ Lex.	5.76	20.47	16.01	39.62
+ PPD	4.98	19.53	13.34	36.49
+ Lex. + PPD	6.36	22.49	16.42	40.04
Seediq	1.52	13.24	8.78	28.02
+ Lex.	2.74	15.69	10.64	30.92
+ PPD	1.79	14.51	10.12	28.44
+ Lex. + PPD	2.91	15.96	11.51	31.04

Table 9: Translation quality achieved using different combinations of data during training, where the first row for each language pair indicates the quality achieved using only the original parallel sentences, +Lex indicates adding bilingual lexicon, and +PPD indicates adding pseudo-parallel data

5.2 Using an externally created lexicon

	to Mandarin		from Mandarin	
Language	BLEU	CHRF	BLEU	CHRF
Seediq	1.52	13.24	8.78	28.02
+ Lex.	1.89	14.24	9.46	27.54
+ PPD	1.68	13.71	8.91	30.49
+ Lex. + PPD	2.03	15.92	9.97	29.63

Table 10: Translation quality achieved after adding an externally created lexicon to the parallel data during training

Though the overlap between the lexicon entries and parallel sentences is also important in evaluating the impact of lexical resources on translation quality, this overlap was difficult to define and examine. For example, in Mandarin Chinese, the definition of a word is vague as nearly each Chinese character is morphologically significant with its own meaning, so the overlap can be defined in multiple ways. Additionally, due to the low-resource nature of the language pairs examined in this paper, we could not find many additional lexicons with which to experiment. We did, however, find one externally created lexicon for Seediq-Mandarin, one with much fewer entries that could also be found in our parallel sentences, to use in our experiments.

To see how our techniques work with an externally created lexicon, one that was created separately from the parallel data, we used a Seediq-Mandarin lexicon of 1,556 entries described in 3. Table 10 displays the translation quality achieved by adding this lexicon and pseudo-parallel data using this lexicon. By adding this lexicon and pseudo-parallel data made using this lexicon to the original parallel data during the training process, we achieved increases of 0.51 BLEU points and 2.68 CHRF points when translating from Seediq to Mandarin and increases of 1.19 BLEU points and 1.61 CHRF points when translating from Mandarin to Seediq. Though these gains are not as great as the translation quality achieved by using the original lexicon, these results still demonstrate that our techniques work. We believe the smaller gains may be attributable to the smaller size of this external lexicon, which contains 1,556 entries which compares unfavorably to ours with 6,723 entries. Though less than half of the Formosan words in this externally lexicon could be found in the parallel sentences, unlike our lexicon, whose Formosan entries nearly all showed up in our parallel sentences, its addition to the training data still produced gains in translation quality. More studies, however, are needed to more clearly define the overlap between lexicons and parallel sentences, especially for Mandarin Chinese, and examine its effect on how adding lexical data can improve translation quality.

5.3 Non-Asian language pair

Finally, to explore how our techniques work with a non-Asian language pair, we considered translation between Spanish and Nahuatl. Adding a lexicon during the training process gave us gains of 0.79 BLEU points and 5.18 CHRF points when translating from Nahuatl to Spanish and gains of 1.48 BLEU points and 11.86 CHRF points when translating from Spanish to Nahuatl. Adding pseudo-parallel data generated with a 20% rate of replacement in addition to this lexicon gave us gains of 1.03 BLEU points and 6.67 CHRF points when translating from Nahuatl to Spanish and gains of 1.98 BLEU points and 21.89 CHRF points when translating from Spanish to Nahuatl. For comparison, the best score achieved in the AmericasNLP 2023 shared task (Ebrahimi et al., 2023) on translation from Spanish to Nahuatl using the same dataset was a BLEU score of 2.33 and a CHRF score of 27.25. This was achieved through extending, training, and combining various adaptations of NLLB-200 (Costa-jussà et al., 2022) and utilizing additional data from other sources including constitutions, handbooks, news articles, and back-translations produced from monolingual data (Gow-Smith and Sánchez Villegas, 2023). Though we did not achieve as high of a BLEU score, we achieved a slightly higher CHRF score using our simple method of using a lexicon and pseudo-parallel data created from it during training.

	to Spanish		from Spanish	
Language	**BLEU**	**CHRF**	**BLEU**	**CHRF**
Nahuatl	0.93	4.25	0.04	5.63
+ Lex.	1.72	9.43	1.52	17.49
+ PPD	1.41	6.94	1.84	24.39
+ Lex. + PPD	1.96	10.92	2.02	27.52

Table 11: Spanish-Nahuatl translation quality

5.4 Multilingual models

Finally, we also explored how models would perform if pretrained on all the Formosan data available to us. Though the Formosan languages are linguistically quite diverse, they are still part of the same language family, and we wanted to see if there were any possibilities for the model to learn from these languages all together. Table 12 displays the translation quality achieved by these multilingual models. On average, the BLEU scores from the model pretrained on all the Formosan languages were 0.15 higher, and the CHRF scores were 0.17 points lower than the models that were not pretrained on anything when finetuning using only parallel sentences. When finetuning using both parallel sentences and the lexicon, the BLEU scores from the model pretrained on all the Formosan languages were 0.07 higher, and the CHRF scores were 0.22 points lower than the models that

	to Mandarin				from Mandarin			
	Parallel sentences		+ Lex.		Parallel sentences		+ Lex.	
Language	BLEU	CHRF	BLEU	CHRF	BLEU	CHRF	BLEU	CHRF
Amis	3.14	11.49	11.03	32.563	11.59	32.524	15.29	39.234
Atayal	3.23	13.519	6.22	13.325	14.53	36.093	14.62	34.497
Bunun	5.3	17.962	6.6	22.98	18.06	40.729	21.07	49.76
Paiwan	1.87	2.89	11.87	12.963	1.2	20.765	9.21	33.598
Sakizaya	3.68	12.705	6.45	19.404	13.13	35.502	14.92	38.538
Kavalan	8.45	24.393	7.8	29.354	28.98	51.824	35.06	57.066
Rukai	9.72	38.353	11.19	39.177	1.02	13.229	2.22	15.626
Puyuma	8.12	15.948	13.66	21.182	18.28	44.069	22.12	48.1
Seediq	1.18	14.093	3.59	15.741	9.02	27.948	11.69	31.827
Thao	9.43	25.761	16.02	31.897	28.37	51.039	30.83	57.14
Saaroa	5.74	14.623	9.02	14.972	8.28	32.85	8.31	40.098
Yami	3.35	17.2	5.54	24.041	21.88	37.19	22.94	44.197
Truku	1.14	6.493	1.36	10.018	9.04	24.014	10.88	31.01
Tsou	0.72	18.41	4.81	23.413	13.86	34.295	17.51	41.628
Kanakanavu	11.27	19.234	12.64	28.23	19.74	43.294	25.21	51.631
Saisiyat	3.34	14.872	7.96	24.486	20.59	43.231	24.84	49.656

Table 12: Translation quality achieved by models that were pretrained on all the Formosan data available and then finetuned for translation for each language pair on just parallel sentences (the “Parallel sentences” columns) and parallel sentences with the lexicon (the “+ Lex.” columns)

were not pretrained on anything on average.

The BLEU scores indicated minimal improvements in quality, and the CHRF scores indicated small decreases in quality. Thus, using multilingual models did not provide any conclusive improvements. This may be due to the fact that the Formosan languages are, in fact, too diverse for the model to learn any meaningful linguistic features from other languages that would help in translation for any given language pair. Additionally, we think our results suggest that the amount of Formosan language data available was also simply too little for cross-lingual pretraining to be meaningful.

6 Conclusion

About 70% of the approximately 7,000 languages of the world have data in the form of lexicons (Wang et al., 2022), and these data are an incredibly valuable resource that can be used to improve low-resource machine translation. When parallel sentence data is limited for languages, it may be difficult for models to learn the translations of different words, and bilingual lexicons can be used to more explicitly teach which words correspond to one another.

By adding lexical data during the training process, we achieved an average gain of 3.37 in BLEU scores and 6.06 in CHRF scores for translation between Mandarin and 16 Formosan languages, whose pairings can all be considered low-resource. These Formosan languages demonstrate considerable variety in linguistic features and are considered the most diverse set of Austronesian languages (Li, 2008). By using existing lexicons and adding them to parallel data, we demonstrate how existing bilingual lexicons, a relatively inexpensive source of data, can be added as parallel data to improve translation quality for low-resource language pairs. We also show that adding more lexical data to parallel data during training tends to lead to higher translation quality. Though adding lexical data is not as effective as adding more parallel data, lexical data are a much cheaper source of data that still generate significant improvements to model output.

When using these lexicons to create pseudo-parallel data, and using both the lexicon entries and this pseudo-parallel data during the training process, we achieved an average increase of 5.55 in BLEU scores and 10.33 in CHRF scores. Adding lexical data was more effective than adding pseudo-parallel data, but adding both lexical data and pseudo-parallel data produced the most gains in translation quality. Additionally, we demonstrate that these techniques also improve translation quality for one non-Asian low-resource language pair, Spanish-Nahuatl. Motivated by these results, we intend to continue exploring how these techniques work for other low-resource language pairs in the world. In the future, we’d also like to have humans evaluate our model outputs, though this remains a challenge due to the low-resource nature of the languages involved.

Limitations

Because the languages that we work with in this paper are extremely low-resource, it was not feasible to find native speakers of the 16 Formosan languages to check the output of our models. The significance of the BLEU and CHRF gains may be more clear by focusing on a smaller set of language pairs to make finding native speakers more manageable and describing the issues in translation output with the help of these native speakers.

We also wanted to explore how externally created lexicons work with these techniques for more language pairs other than Seediq-Mandarin, but we were not able to find any additional bilingual lexicons for other language pairs. This may be resolved by examining paper resources that may not yet be available digitally.

Additionally, we cannot ascertain how well our results generalize to other languages. Formosan languages have a considerable amount of linguistic variation, and our techniques work for Spanish-Nahuatl translation, but we were unable to experiment with and demonstrate results for other language pairs.

References

R.A. Blust. 2003. *Thao Dictionary*. Language and linguistics monograph. Series A. Institute of Linguistics (Preparatory Office), Academic Sinica.

Robert Blust. 2013. *The Austronesian languages*. Asia-Pacific Linguistics, School of Culture, History and Language, College of Asia and the Pacific, The Australian National University.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

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*.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219, Toronto, Canada. Association for Computational Linguistics.

Aloka Fernando, Surangika Ranathunga, Dilan Sachintha, Lakmali Piyarathna, and Charith Rajitha. 2023. Exploiting bilingual lexicons to improve multilingual embedding-based document and sentence alignment for low-resource languages. *Knowledge and Information Systems*, 65(2):571–612.

Edward Gow-Smith and Danae Sánchez Villegas. 2023. Sheffield's submission to the AmericasNLP shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 192–199, Toronto, Canada. Association for Computational Linguistics.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016. Axolotl: a web accessible parallel corpus for Spanish-Nahuatl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4210–4214, Portorož, Slovenia. European Language Resources Association (ELRA).

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Aizhan Imankulova, Takayuki Sato, and Mamoru Komachi. 2017. Improving low-resource neural machine translation with filtered pseudo-parallel corpus. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 70–78, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Alex Jones, Isaac Caswell, and Orhan Firat. 2023. BiLex Rx: Lexical data augmentation for massively multilingual machine translation.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3145–3150, Reykjavik, Iceland. European Language Resources Association (ELRA).

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Jonathan Cheng-Chuen Kuo. 2015. *Argument alternation and argument structure in symmetrical voice languages: A case study of transfer verbs in Amis, Puyuma, and Seediq*. Ph.D. thesis, University of Hawai'i at Manoa.

Amy Pei-jung Lee et al. 2011. Comitative and coordinate constructions in truku seediq. *Language and linguistics*, 12(1):49–75.

Paul Jen-kuei Li. 2008. The great diversity of Formosan languages. *Language and Linguistics*, 9(3):523–546.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663, Online. Association for Computational Linguistics.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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

Machel Reid, Junjie Hu, Graham Neubig, and Yutaka Matsuo. 2021. AfroMT: Pretraining strategies and reproducible benchmarks for translation of 8 African languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1306–1320, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Shigeru Tsuchida. 1990. Classificatory prefixes of tsou verbs. *Tokyo University Linguistics Papers*, 89:17–52.

Lonneke van der Plas, Paola Merlo, and James Henderson. 2011. Scaling up automatic cross-lingual semantic role annotation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 299–304, Portland, Oregon, USA. Association for Computational Linguistics.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877, Dublin, Ireland. Association for Computational Linguistics.

A Train-Dev-Test Split

	Train	Dev	Test
Amis	4,600	576	575
Atayal	4,600	576	575
Bunun	7,180	898	897
Kanakanavu	5,294	662	662
Kavalan	6,573	822	821
Paiwan	4,126	516	516
Puyuma	5,515	689	690
Rukai	8,319	1,040	1,040
Saaroa	3,839	480	480
Saisiyat	4,839	605	605
Sakizaya	4,590	574	573
Seediq	4,367	546	546
Thao	5,952	744	744
Truku	3,678	460	459
Tsou	3,550	444	443
Yami	5,186	648	649

Table 13: Number of parallel sentences with Mandarin in the train, dev, and test sets