Towards Computational Linguistics in Minangkabau Language: Studies on Sentiment Analysis and Machine Translation

Fajri Koto
School of Computing and Information System
The University of Melbourne
ffajri@student.unimelb.edu.au

Ikhwan Koto
Faculty of IT
Andalas University
ikhwan220397@gmail.com

Abstract

Although some linguists (Rusmali et al., 1985; Crouch, 2009) have fairly attempted to define the morphology and syntax of Minangkabau, information processing in this language is still absent due to the scarcity of the annotated resource. In this work, we release two Minangkabau corpora: sentiment analysis and machine translation that are harvested and constructed from Twitter and Wikipedia.[1] We conduct the first computational linguistics in Minangkabau language employing classic machine learning and sequence-to-sequence models such as LSTM and Transformer. Our first experiments show that the classification performance over Minangkabau text significantly drops when tested with the model trained in Indonesian. Whereas, in the machine translation experiment, a simple word-to-word translation using a bilingual dictionary outperforms LSTM and Transformer model in terms of BLEU score.

1 Introduction

Minangkabau (Baso Minang) is an Austronesian language with roughly 7m speakers in the world (Gordon, 2005). The language is spread under the umbrella of Minangkabau tribe – a matrilineal culture in the province of West Sumatra, Indonesia. The first-language speakers of Minangkabau are scattered across Indonesian archipelago and Negeri Sembilan, Malaysia due to "*merantau*" (migration) culture of Minangkabau tribe (Drakard, 1999).

Despite there being over 7m first-language speakers of Minangkabau,[2] this language is rarely used in the formal sectors such as government and education. This is because the notion to use Bahasa Indonesia as the unity language since the Independent day of Indonesia in 1945 has been a double-edged sword. Today, Bahasa Indonesia successfully connects all ethnicities across provinces in Indonesia (Cohn et al., 2014), yet threatens the existent of some indigenous languages as the native speakers have been gradually decreasing (Novitasari et al., 2020). Cohn et al. (2014) predicted that Indonesia may shift into a monolingual society in the future.

In this paper, we initiate the preservation and the first information processing of Minangkabau language by constructing a Minangkabau–Indonesian parallel corpus, sourced from Twitter and Wikipedia. Unlike other indigenous languages such as Javanese and Sundanese that have been discussed in machine speech chain (Novitasari et al., 2020; Wibawa et al., 2018), part-of-speech (Pratama et al., 2020), and translation system (Suryani et al., 2016), information processing in Minangkabau language is less studied. To the best of our knowledge, this is the first research on NLP in Minangkabau language, which we conduct in two different representative NLP tasks: sentiment analysis (classification) and machine translation (generation).

There are two underlying reasons why we limit our work in Minangkabau–Indonesian language pair. First, Minangkabau and Indonesian language

[1] Our data can be accessed at https://github.com/fajri91/minangNLP

[2] In Indonesia, Minangkabau language is the fifth most spoken indigenous language after Javanese (75m), Sundanese (27m), Malay (20m), and Madurese (14m) (Riza, 2008).

are generally intelligible with some overlaps of lexicons and syntax. The Indonesian language has been extensively studied and is arguably a convenient proxy to learn the Minangkabau language. Second, authors of this work are the first-language speakers of Minangkabau and Indonesian language. This arguably eases and solidifies the research validation in both tasks.

To summarize, our contributions are: (1) we create a bilingual dictionary from Minangkabau Wikipedia by manually translating top 20,000 words into Indonesian; 2) we release Minangkabau corpus for sentiment analysis by manually translating 5,000 sentences of Indonesian sentiment analysis corpora; 3) We develop benchmark models with classic machine learning and pre-trained language model for Minangkabau sentiment analysis; 4) We automatically create a high-quality machine translation corpus consisting 16K Minangkabau–Indonesian parallel sentences; and 5) We showcase the first Minangkabau–Indonesian translation system through LSTM and Transformer model.

2 Minangkabau–Indonesian Bilingual Dictionary

In the province of West Sumatra, Minangkabau language is mostly used in spoken communication, while almost all reading materials such as local newspaper and books are written in Indonesian. Interestingly, Minangkabau language is frequently used in social media such as Twitter, Facebook and WhatsApp, although the writing can be varied and depends on the speaker dialect. Rusmali et al. (1985) define 6 Minangkabau dialects based on cities/regencies in the West Sumatra province. This includes Agam, Lima Puluh Kota, Pariaman, Tanah Datar, Pesisir Selatan, and Solok. The variation among these dialects is mostly phonetic and rarely syntactic.

Crouch (2009) classifies Minangkabau language into two types: 1) Standard Minangkabau and 2) Colloquial Minangkabau. The first type is the standard form for intergroup communication in the province of West Sumatra, while the second is the dialectal variation and used in informal and familiar contexts. Moussay (1998) and Crouch (2009) argue that Padang dialect is the standard form of Minangkabau. However, as the first-language speaker, we contend that these statements are inaccurate because of two reasons. First, many locals do not aware of Padang dialect. We randomly survey 28 local people and only half of them know the existence of Padang dialect. Second, in 2015 there has been an attempt to standardize Minangkabau language by local linguists, and Agam-Tanah Datar is proposed as the standard form due to its largest population.[3]

Our first attempt in this work is to create a publicly available Minangkabau–Indonesian dictionary by utilising Wikipedia. Minangkabau Wikipedia[4] has 224,180 articles (rank 43rd) and contains 121,923 unique words, written in different dialects. We select top-20,000 words and manually translate it into Indonesian. We found that this collection contains many noises (e.g. scientific terms, such as *onthophagus*, *molophilus*) that are not Minangkabau nor Indonesian language. After manually translating the Minangkabau words, we use *Kamus Besar Bahasa Indonesia* (KBBI)[5] – the official dictionary of Indonesian language to discard the word pairs with the unregistered Indonesian translation. We finally obtain 11,905-size Minangkabau–Indonesian bilingual dictionary, that is 25 times larger than word collection in Glosbe (476 words).[6]

We found that 6,541 (54.9%) Minangkabau words in the bilingual dictionary are the same with the translation. As both Minangkabau and Indonesian languages are Austronesian (*Malayic*) language, the high ratio of lexicon overlap is very likely. Further, we observe that 1,762 Indonesian words have some Minangkabau translations. These are primarily synonyms and dialectal variation that we show in Table 1. Next, in this study, we use this dictionary in sentiment analysis and machine translation.

3 Sentiment Analysis

Sentiment analysis has been extensively studied in English and Indonesia in different domains such as movie review (Yessenov and Misailovic, 2009; Nurdiansyah et al., 2018), Twitter (Agarwal et al., 2011;

[3] https://id.wikimedia.org/wiki/Sarasehan_Bahasa_Minangkabau

[4] Downloaded in June 2020

[5] https://github.com/geovedi/indonesian-wordlist

[6] https://glosbe.com/min/id

Indonesian	English	Minangkabau
	Synonyms	
ibunya	her mother	*ibunyo, mandehnyo, amaknyo*
memperlihatkan	to show	*mampacaliak, mampaliekan*
kelapa	coconut	*karambia, kalapo*
	Dialectal variations	
berupa	such as	*barupo, berupo, berupa, barupa*
bersifat	is, act, to have the quality	*basipaik, basifaik, basifek, basifat, basipek*
Belanda	Netherlands	*Balando, Belanda, Bulando, Belando*

Table 1: Example of synonyms and dialectal variations in the Minangkabau–Indonesian dictionary

Koto and Adriani, 2015), and presidential election (Wang et al., 2012; Ibrahim et al., 2015). It covers a wide range of approaches, from classic machine learning such as naive Bayes (Nurdiansyah et al., 2018), SVM (Koto and Adriani, 2015) to pre-trained language models (Sun et al., 2019; Xu et al., 2019). The task is not only limited to binary classification of positive and negative polarity, but also multi classification (Liu and Chen, 2015), subjectivity classification (Liu, 2010), and aspect-based sentiment (Ma et al., 2017).

In this work, we conduct a binary sentiment classification on positive and negative sentences by first manually translating Indonesian sentiment analysis corpus to Minangkabau language (Agam-Tanah Datar dialect). To provide a comprehensive preliminary study, we experimented with a wide range of techniques, starting from classic machine learning algorithms, recurrent models, to the state of the art technique, Transformer (Vaswani et al., 2017).

3.1 Dataset

The data we use in this work is sourced from 1) Koto and Rahmaningtyas (2017); and 2) an aspect-based sentiment corpus.[7] Koto and Rahmaningtyas (2017) dataset is originally from Indonesian tweets and has been labelled with positive and negative class. The second dataset is a hotel review collection where each review can encompass multi-polarity on different aspects. We determine the sentiment class based on the majority count of the sentiment label, and simply discard it if there is a tie between positive and negative. In total, we obtain 5,000 Indonesian texts from these two sources. We then ask two native speakers of Minangkabau and Indonesian language to manually translate all texts in the corpus. Finally, we create a parallel sentiment analysis corpus with 1,481 positive and 3,519 negative labels.

[7] https://github.com/annisanurulazhar/absa-playground/

3.2 Experimental Setup

We conducted two types of the zero-shot experiment by using Indonesian train and development sets. In the first experiment, the model is tested against Minangkabau data, while in the second experiment we test the same model against the Indonesian translation, obtained by word-to-word translation using the bilingual dictionary (Section 2). There are two underlying reasons to perform the zero-shot learning: 1) Minangkabau is intelligible with Indonesian language and most available corpus in the West Sumatra is Indonesian; 2) Minangkabau language is often mixed in Indonesian data collection especially in social media (e.g. Twitter, if the collection is based on geographical filter). Through zero-shot learning, we aim to measure the performance drop of Indonesian model when tested against the indigenous language like Minangkabau.

Our experiments in this section are based on 5-folds cross-validation. We conduct stratified sampling with ratio 70/10/20 for train, development, and test respectively, and utilize five different algorithms as shown in Table 2. For naive Bayes, SVM and logistic regression, we use byte-pair encoding (unigram and bigram) during the training and tune the model based on the development set. Due to data imbalance, we report the averaged F-1 score of five test sets.

For Bi-LSTM (200-d hidden size) we use two

Method	Train $_{ID}$		Train $_{MIN}$
	Test $_{MIN}$	Test $_{ID'}$	Test $_{MIN}$
Naive Bayes	68.49	68.86	73.03
SVM	59.75	68.35	74.05
Logistic Regression	57.95	66.90	72.35
Bi-LSTM	58.75	65.62	72.37
Bi-LSTM + `fastText`	62.06	71.51	70.47
MBERT	62.71	67.60	**75.91**

Table 2: Results for Sentiment Analysis on Minangkabau test set. The numbers are the averaged F-1 of 5-folds cross validation sets. MIN = Minangkabau, ID = Indonesian, ID' = Indonesian translation through bilingual dictionary.

variants of 300-d word embedding: 1) random initialization; and 2) `fastText` pre-trained Indonesian embeddings (Bojanowski et al., 2016). First, we lowercase all characters and truncate them by 150 maximum words. We use batch size 100, and concatenate the last hidden states of Bi-LSTM for classification layer. For each fold, we train and tune the model for 100 steps with Adam optimizer and early stopping (patience = 20).

Lastly, we incorporate the Transformer-based language model BERT (Devlin et al., 2019) in our experiment. Multilingual BERT (MBERT) is a masked language model trained by concatenating 104 languages in Wikipedia, including Minangkabau. MBERT has been shown to be effective for zero-shot cross-lingual tasks including classification (Wu and Dredze, 2019). In this work, we show the first utility of MBERT for classifying text in the indigenous language, such as Minangkabau. In fine-tuning, we truncate all data by 200 maximum tokens, and use batch size 30 and maximum epoch 20 (2,500 steps). The initial learning rate is 5e-5 with warm-up of 10% of the total steps. We evaluate F-1 score of the development set for every epoch, and terminate the training if the performance does not increase within 5 epochs. Similar to Bi-LSTM models, we use Adam optimizer for gradient descent steps.

3.3 Result

In Table 2, we show three different experimental results. The first column is the zero-shot setting where the model is trained and tuned using Indonesian text and tested against Minangkabau data. Surprisingly, naive Bayes outperforms other models including MBERT with a wide margin. Naive Bayes achieves 68.49 F1-score, +6 points over the pre-trained language model and Bi-LSTM + `fastText`. This might indicate that naive Bayes can effectively exploit the vocabulary overlap between Minangkabau and Indonesian language.

In the second experiment, we hypothesize that a simple word-to-word translation using a bilingual dictionary can improve zero-shot learning. Similar to the first experiment, we train the model with Indonesian text, but we test the model against the Indonesian translation. As expected, the F-1 scores improve dramatically for all methods except naive Bayes with +0.37 gains. SVM, logistic regression and Bi-LSTMs are improved by 6–9 points while MBERT gains by +5 points by predicting the Indonesian translation.

In the third experiment, we again show a dramatic improvement when the model is fully trained in the Minangkabau language. Compared to the second experiment, all models are improved by 4–8 points with Bi-LSTM + `fastText` in exception. This is because the model uses `fastText` pre-trained Indonesian embeddings, and its best utility is when the model is trained and tested in the Indonesian language (second experiment). The best model is achieved by MBERT with 75.91 F1-score, outperforming other models with a comfortable margin.

Based on these experiments, we can conclude the necessity of specific indigenous language resource for text classification in Indonesia. These languages are mixed in Indonesian social media, and testing the Indonesian model directly on this Indonesian-type language can drop the sentiment classification performance by 11.41 on average.

3.4 Error Analysis

In this section, we manually analyze the false positive (FN) and false negative (FP) of MBERT model. We examine all misclassified instances in the test set by considering three factors:

- *Bias towards a certain topic*. In Indonesia, we argue that public sentiment towards government, politics and some celebrities are often negative. This could lead to bias in the train-

Category	Value
#FN	83
Bias towards certain topic (%)	34.84
Single polarity with negative words (%)	20.48
Mixed polarity (%)	12.05
#FP	56
Bias towards certain topic (%)	26.79
Single polarity with positive words (%)	26.79
Mixed polarity (%)	28.57

Table 3: Error analysis for False Negative (FN) and False Positive (FP) set.

ing and result in a wrong prediction in the test set. We count the number of texts in FP and FN set that contain these two topics: politics and celebrity.

- *Single polarity but containing words in opposite polarity*. The model might fail to correctly predict a sentiment label when contains words with the opposite polarity.
- *Mixed polarity*. A text can consist of both positive and negative polarity with one of them is more dominant.

In Table 3 we found that there are 83 FN (28% of positive data) and 56 FP (8% of negative data) instances. We further observe that 34.84% of FN instances contain politics or celebrity topic, while there is only 26.79% of FP instances with these criteria. In Figure 1, we show an FN example for the first factor: "*Iduik Golkar! Idrus jo Yorrys Bapaluak*" where "*Golkar*" is one of the political parties in Indonesia.

Secondly, 20.48% of FN instances contain negative words. As shown in Figure 1 the example uses words "give up" and "hurt" to convey positive advice. We notice that the second factor is more frequent in FP instances with 26.79% proportion. Lastly, we find that 28.57% of FP instances have mixed polarity, 2 times larger than FN. We observe that most samples with mixed polarity are sourced from the hotel review. It highlights that mixed polarity is arguably a harder task, and requires special attention to aspects in text fragments.

4 Machine Translation

Machine translation has been long run research, started by Rule-based Machine Translation (RBMT) (Carbonell et al., 1978; Nagao, 1984), Statistical Machine Translation (SMT) (Brown et al., 1990), to Neural Machine Translation (NMT) (Bahdanau et al., 2015). NMT with its continuous vector representation has been a breakthrough in machine translation, minimizing the complexity of SMT yet boosts the BLEU score (Papineni et al., 2002) into a new level. Recently, Hassan et al. (2018) announce that their Chinese–English NMT system has achieved a comparable result with human performance.

Although there are 2,300 languages across Asia, only some Asian languages such as Chinese and Japanese have been extensively studied for machine translation. We argue there are two root causes: a lack of parallel corpus, and a lack of resource standardization. Apart from the Chinese language, there have been some attempts to create a parallel corpus across Asian languages. Nomoto et al. (2019) construct 1,3k parallel sentences for Japanese, Burmese, Malay, Indonesian, Thai, and Vietnamese, while Kunchukuttan et al. (2017) release a large-scale Hindi-English corpus. Unlike these national languages, machine translation on indigenous languages is still very rare due to data unavailability. In Indonesia, Sundanese (Suryani et al., 2015) and Javanese (Wibawa et al., 2013) have been explored through statistical machine translation. In this work, our focus is Minangkabau–Indonesian language pair, and we first construct the translation corpus from Wikipedia.

4.1 Dataset

Constructing parallel corpus through sentence alignment from bilingual sources such as news (Zhao and Vogel, 2002; Rauf and Schwenk, 2011), patent (Utiyama and Isahara, 2003; Lu et al., 2010), and Wikipedia (Yasuda and Sumita, 2008; Smith et al., 2010; Chu et al., 2014) have been done in various language pairs. For Indonesian indigenous language, Trisedya and Inastra (2014) has attempted to create parallel Javanese–Indonesian corpus by utilizing inter-Wiki links and aligning sentences via Gale and Church (1993) approach.

In this work, we create Minangkabau–Indonesian

Minangkabau	English
Bias towards certain topic: Iduik Golkar! Idrus jo Yorrys Bapaluak	**Bias towards certain topic:** Glory for Golkar! Idrus and Yorrys are hugging
Positive text that uses negative words: Ado masonyo dima awak harus marelaan nan awak sayangi untuak pai.. Yo mamang sakik! Tapi tu demi kebaikan awak surang. :)	**Positive text that uses negative words:** there are times when we have to give up on someone we care about. Yes indeed hurt! but this is for our good :)
Mixed polarity: kamarnyo rancak, patamu kali pakai airy kironyo dapek toolkit dan snack lo, cuma wifi hotelnyo indak bisa connect.. overall rancak, apolai diskon 80%	**Mixed polarity:** the room was good, the first time I used Airy, it turns out I got snacks and toolkits, it's just that hotel wifi was not functioning. Overall is good, especially 80% discounts.

Figure 1: Example of False Negative.

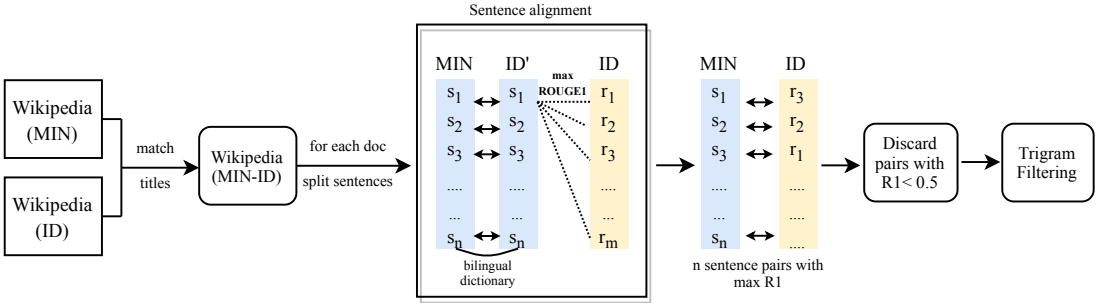

Figure 2: Flow chart of MIN-ID parallel corpus construction.

(MIN-ID) parallel corpus by using Wikipedia[8] (Figure 2). We obtain 224,180 Minangkabau and 510,258 Indonesian articles, and align documents through title matching, resulting in 111,430 MIN-ID document pairs. After that, we do sentence segmentation based on simple punctuation heuristics and obtain 4,323,315 Minangkabau sentences. We then use the bilingual dictionary (Section 2) to translate Minangkabau article (MIN) into Indonesian language (ID'). Sentence alignment is conducted using ROUGE-1 (F1) score (unigram overlap) (Lin, 2004) between ID' and ID, and we pair each MIN sentence with an ID sentence based on the highest ROUGE-1. We then discard sentence pairs with a score of less than 0.5 to result in 345,146 MIN-ID parallel sentences.

We observe that the sentence pattern in the collection is highly repetitive (e.g. 100k sentences are about biological term definition). Therefore, we conduct final filtering based on top-1000 trigram by iteratively discarding sentences until the frequency of each trigram equals to 100. Finally, we obtain 16,371 MIN-ID parallel sentences and conducted manual evaluation by asking two native Minangkabau speakers to assess the adequacy and fluency (Koehn and Monz, 2006). The human judgement is based on scale 1–5 (1 means poor quality and 5 otherwise) and conducted against 100 random samples. We average the weights of two annotators before computing the overall score, and we achieve 4.98 and 4.87 for adequacy and fluency respectively.[9] This indicates that the resulting corpus is high-quality for machine translation training.

[8]Downloaded in June 2020

Category	Wiki		SentC	
	MIN	ID	MIN	ID
mean(#word)	19.6	19.6	22.3	22.2
std(#word)	11.6	11.5	12.8	12.7
mean(#char)	105.2	107.7	98.9	99.2
std(#char)	59.1	60.4	57.9	58.1
#vocab	32,420	27,318	13,940	13,698
Overlapping #vocab	21,563		9,508	

Table 4: Statistics of machine translation corpora.

4.2 Experimental Setup

First, we split Wikipedia data with ratio 70/10/20, resulting in 11,571/1,600/3,200 data for train, de-

[9]The Pearson correlation of two annotators for adequacy and fluency are 0.9433 and 0.5812 respectively

velopment, and test respectively. In addition, we use parallel sentiment analysis corpus (Section 3) as the second test set (size 5,000) for evaluating texts from different domain. In Table 4, we provide the overall statistics of both corpora: Wikipedia (Wiki) and Sentiment Corpus (SentC). We observe that Minangkabau (MIN) and Indonesian (ID) language generally have similar word and char lengths. The difference is in the vocabulary size where Minangkabau is 5k larger than Indonesian in Wiki corpus. As we discuss in Section 2, this difference is due to various Minangkabau dialects in Wikipedia.

We conducted two experiments: 1) Minangkabau to Indonesian (MIN → ID); and 2) Indonesian to Minangkabau (ID → MIN) with three models: 1) word-to-word translation (W2W) using bilingual dictionary (Section 2); 2) LSTMs; and 3) Transformer. We use Moses Tokeniser[10] for tokenization, and sacreBLEU script (Post, 2018) to evaluate BLEU score on the test sets. All source and target sentences are truncated by 75 maximum lengths.

Our encoder-decoder (LSTM and Transformer) models are based on Open-NMT implementation (Klein et al., 2017). For LSTM models, we use two layers of 200-d Bi-LSTM encoder and 200-d LSTM decoder with a global attention mechanism. Source and target embeddings are 500-d and shared between encoder and decoder. For training, we set the learning rate of 0.001 with Adam optimizer, and warm-up of 10% of the total steps. We train the model with batch size 64 for 50,000 steps and evaluate the development set for every 5,000 steps.

The Transformer encoder-decoder (each) has 6 hidden layers, 512 dimensionality, 8 attention heads, and 2,028 feed-forward dimensionalities. Similar to LSTM model, the word embeddings are shared between source and target text. We use cosine positional embedding and train the model with batch size 5,000 for 50,000 steps with Adam optimizer (warm-up = 5,000 and Noam decay scheme). We evaluate the development set for every 10,000 steps.

4.3 Result

In Table 5, we present the experiment results for machine translation. Because Indonesian and Minangkabau language is mutually intelligible, and Table 4 shows that roughly 75% words in two vocabularies overlap, we set the BLEU scores of raw source and target text as the baseline. We found for both MIN→ID and ID→MIN, the BLEU scores are relatively high, more than 30 points.

Method	MIN → ID		ID → MIN	
	Wiki	SentC	Wiki	SentC
Raw (baseline)	30.08	43.73	30.08	43.73
W2W	**64.54**	**60.99**	**55.08**	**55.22**
LSTM	63.77	22.82	48.50	15.52
Transformer	56.25	10.23	43.50	8.86

Table 5: BLEU score on the test set. SentC is parallel sentiment analysis corpus in Section 3.

We observe that a simple word-to-word (W2W) translation using a bilingual MIN-ID dictionary achieves the best performance over LSTM and Transformer model in all cases. For MIN→ID, the BLEU scores are 64.54 and 60.99 for Wiki and SentC respectively, improving the baseline roughly 20–30 points. The similar result is also found in ID→MIN with disparity 12–25 points in the baseline.

Both LSTM and Transformer models significantly improve the baseline for Wiki corpus, but poorly perform in translating SentC dataset. For the Wiki corpus, the LSTM model achieves a competitive score in MIN→ID and ID→MIN, improving the baseline for 33 and 18 points respectively. The Transformer also outperforms the baselines, but substantially lower than the LSTM. In out-of-domain test set (SentC), the performance of both models significantly drops, 20–30 points lower than the baselines. We further observe that this is primarily due to out of vocabulary issue, where around 65% words in SentC are not in the vocabulary model.

4.4 Analysis

In Figure 3, we show translation examples in Wiki corpus. In MIN→ID, the LSTM translation is slightly more eloquent than word-to-word (W2W) translation. Word "*tamasuak*" (including) in W2W translation is Minangkabau language and not properly translated. This is because the word "*tama-*

[10] https://pypi.org/project/mosestokenizer/

Wiki corpus Example (MIN-ID)	Wiki corpus Example (ID-MIN)
Source (MIN): saketek nan dikatahui tantang kahidupan awal ching shih, tamasuak namo lahia jo tanggal lahirnya (*There is little information about Ching Shih early life, including her birth name and birth date*) **Target (ID):** **Reference:** sedikit yang diketahui tentang kehidupan awal ching shih, termasuk nama lahir dan tanggal lahirnya **W2W with Bilingual Dictionary:** sedikit yang diketahui tantang kehidupan awal ching shih, tamasuak nama lahir dengan tanggal lahirnya **LSTM:** sedikit yang diketahui tentang kehidupan awal ching shih, termasuk nama lahir dan tanggal lahirnya **Transformer:** sedikit yang diketahui tentang kehidupan awal rambut, termasuk nama lahir dan tanggal kelahiran	**Source (ID):** laba-laba ini biasanya banyak ditemui di amerika serikat, guatemala, antigua (*This spider is mostly found in the United States, Guatemala, Antigua*) **Target (MIN):** **Reference:** lawah iko biasonyo banyak ditamui di amerika sarikat, guatemala, antigua **W2W with Bilingual Dictionary:** laba-laba ko biasonyo banyak ditemui di amerika serikat, guatemala, antigua **LSTM:** lawah iko biasonyo banyak ditamui di amerika serikat, guatemala, moldavia **Transformer:** lawah iko biasonyo banyak ditamui di amerika sarikat, guatemala, alaska

Figure 3: Examples of model translation in Wikipedia corpus.

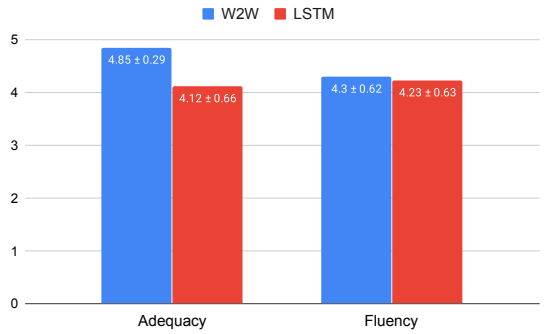

Figure 4: Human judgment on MIN→ID (Wiki)

suak" is not registered in the bilingual vocabulary. In this sample, Transformer model hallucinates as mentioning "*kehidupan awal rambut*" (early life of hair), resulting in a poor fluency and adequacy. Next in ID→MIN, W2W translation is better than LSTM and Transformer. W2W translation is relatively good, despite word "*ditemui*" (found) that is not translated into Minangkabau. In this example, LSTM and Transformer hallucinate mentioning incorrect location such as "*moldavia*", and "*alaska*".

For further analysis, we conduct a manual evaluation on two best models: W2W and LSTM in MIN→ID (Wiki) experiment. Like manual evaluation in Section 4.1, we ask two native Indonesian and Minangkabau speakers to examine the adequacy and fluency of 100 random samples with scale 1–5. Figure 1 shows that W2W translation significantly better than LSTM in terms of adequacy, but similar in terms of fluency. This is in line with our observation, that the LSTM model frequently generates incorrect keywords in a fluent and coherent translation. This is possibly due to the out-of-vocabulary (OOV) case in the test set, triggered by the small-size of our train set. A proper training scheme for the low-resource setting can be leveraged in future work, so it can reduce hallucination issue in the LSTM model.

5 Conclusion

In this work, we have shown the first NLP tasks in Minangkabau language. In sentiment analysis task, we found the necessity of indigenous language corpus for classifying Indonesian texts. Although Indonesian and Minangkabau languages are from *Malayic* family, the Indonesian model can not optimally classify Minangkabau text. Next, in the machine translation experiment, although the word-to-word translation is superior to LSTM and Transformer, there is still a room of improvement for fluency. This can be addressed by training seq-to-seq model with a larger corpus.

Acknowledgments

We are grateful to the anonymous reviewers for their helpful feedback and suggestions. In this research, Fajri Koto is supported by the Australia Awards Scholarship (AAS), funded by Department of Foreign Affairs and Trade (DFAT) Australia.

References

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment Analysis of Twitter Data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*. pages 30–38.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR 2015 : International Conference on Learning Representations 2015*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *A statistical approach to machine translation*, 16(2): 79–85.

Jaime G Carbonell, Richard E Cullinford, and Anatole V Gershman. 1978. *Knowledge-based machine translation.* Technical report, Yale University, Department of Computer Science, Connecticut, US.

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2014. Constructing a Chinese–Japanese Parallel Corpus from Wikipedia. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. pages 642–647.

Abigail C Cohn, and Maya Ravindranath. 2014. Local languages in Indonesia: Language maintenance or language shift. *Linguistik Indonesia*, 32(2): 131–148.

Sophie Elizabeth Crouch. 2009. *Voice and verb morphology in Minangkabau, a language of West Sumatra, Indonesia.* Master Thesis, The University of Western Australia.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT 2019: Annual Conference of the North American Chapter of the Association for Computational Linguistics*. pages 4171–4186.

Jane Drakard. 1999. *A Kingdom of Words: Language and Power in Sumatra.*

William A. Gale, and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistic*, 19(1): 75–102.

Raymond G. Gordon. 2005. *Ethnologue: languages of the world, Fifteenth Edition.* SIL International, Dallas, Texas.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. *arXiv preprint arXiv:1803.05567*.

Mochamad Ibrahim, Omar Abdillah, Alfan F. Wicaksono, and Mirna Adriani. 2015. Buzzer Detection and Sentiment Analysis for Predicting Presidential Election Results in a Twitter Nation. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. pages 1348–1353.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proc. ACL*.

Philipp Koehn and Christof Monz. 2006. Manual and Automatic Evaluation of Machine Translation between European Languages. In *Proceedings on the Workshop on Statistical Machine Translation*. pages 102–121.

Fajri Koto and Mirna Adriani. 2015. A Comparative Study on Twitter Sentiment Analysis: Which Features are Good? In *20th International Conference on Applications of Natural Language to Information Systems, NLDB 2015*. pages 453–457.

Fajri Koto and Mirna Adriani. 2015. The Use of POS Sequence for Analyzing Sentence Pattern in Twitter Sentiment Analysis. In *2015 IEEE 29th International Conference on Advanced Information Networking and Applications Workshops*. pages 547–551.

Fajri Koto and Gemala Y. Rahmaningtyas. 2017. InSet lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs. In *2017 International Conference on Asian Language Processing (IALP)*. pages 391–394.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2017. The IIT Bombay English-Hindi parallel corpus. *arXiv preprint arXiv:1710.02855*.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. pages 74–81.

Shuhua Monica Liu and Jiun-Hung Chen. 2015. A multi-label classification based approach for sentiment classification. *Expert Systems With Applications*, 42(3): 1083–1093

Bing Liu. 2010. Sentiment Analysis and Subjectivity. In *Handbook of Natural Language Processing*. pages 627–666.

Bin Lu, Tao Jiang, Kapo Chow, and Benjamin K. Tsou. 2010. Building a Large English-Chinese Parallel Corpus from Comparable Patents and its Experimental Application to SMT.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *IJCAI'17 Proceedings of the 26th International Joint Conference on Artificial Intelligence*. pages 4068–4074.

Gérard Moussay. 1998. Tata Bahasa Minangkabau. Kepustakaan Populer Gramedia, Jakarta.

Makoto Nagao. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. In *Proc. of the international NATO symposium on Artificial and human intelligence*. pages 173–180.

Hiroki Nomoto, Kenji Okano, Sunisa Wittayapanyanon, and Junta Nomura. 2019. Interpersonal meaning annotation for Asian language corpora: The case of TUFS Asian Language Parallel Corpus (TALPCo). In *Proceedings of the Twenty-Fifth Annual Meeting of the Association for Natural Language Processing*. pages 846–849.

Sashi Novitasari, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2020. Cross-Lingual Machine Speech Chain for Javanese, Sundanese, Balinese, and Bataks Speech Recognition and Synthesis. In *SLTU/CCURL@LREC*. pages 131–138.

Yanuar Nurdiansyah, Saiful Bukhori, and Rahmad Hidayat. 2018. Sentiment Analysis System for Movie Review in Bahasa Indonesia using Naive Bayes Classifier Method. In *Journal of Physics: Conference Series*. pages 12011.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. pages 311–318.

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.

Ryan Armiditya Pratama, Arie Ardiyanti Suryani, and Warih Maharani. 2020. Part of Speech Tagging for Javanese Language with Hidden Markov Model. In *Journal of Computer Science and Informatics Engineering (J-Cosine)*, 4(1): 84–91.

Sadaf Abdul Rauf and Holger Schwenk. 2011. Parallel sentence generation from comparable corpora for improved SMT. In *Machine Translation*, 25(4): 341–375.

Hammam Riza. 2008. Resources Report on Languages of Indonesia. In *ALR@IJCNLP*. pages 93–94.

Marah Rusmali, Amir Hakim Usman, Syahwin Nikelas, Nurzuir Husin, Busri Busri, Agusli Lana, M. Yamin, Isna Sulastri, and Irfani Basri. 1985. Kamus Minangkabau – Indonesia. Pusat Pembinaan dan Pengembangan Bahasa Departemen Pendidikan dan Kebudayaan, Jakarta.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. pages 403–411.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. In *NAACL-HLT*. pages 380–385.

Arie Ardiyanti Suryani, Dwi Hendratmo Widyantoro, Ayu Purwarianti, and Yayat Sudaryat. 2015. Experiment on a phrase-based statistical machine translation using POS Tag information for Sundanese into Indonesian. In *2015 International Conference on Information Technology Systems and Innovation (ICITSI)*. pages 1–6.

Arie Ardiyanti Suryani, Isye Arieshanti, Banu W. Yohanes, M. Subair, Sari D. Budiwati, and Bagus S. Rintyarna. 2016. Enriching English into Sundanese and Javanese translation list using pivot language. In *2016 International Conference on Information & Communication Technology and Systems (ICTS)*. pages 167–171.

Bayu Distiawan Trisedya and Dyah Inastra. 2014. Creating Indonesian-Javanese parallel corpora using wikipedia articles. In *2014 International Conference on Advanced Computer Science and Information System*. pages 239–245.

Masao Utiyama and Hitoshi Isahara. 2003. Reliable Measures for Aligning Japanese-English News Articles and Sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. pages 72–79.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. pages 5998–6008.

Hao Wang, Dogan Can, Abe Kazemzadeh, Franois Bar, and Shrikanth Narayanan. 2012. A System for Realtime Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle. In *Proceedings of the ACL 2012 System Demonstrations*. pages 115–120.

Aji P. Wibawa, Andrew Nafalski, A. Effendi Kadarisman, and Wayan F. Mahmudy. 2013. Indonesian-to-Javanese Machine Translation. In *International journal of innovation, management and technology*.

Jaka Aris Eko Wibawa, Supheakmungkol Sarin, Chen Fang Li, Knot Pipatsrisawat, Keshan Sodimana, Oddur

Kjartansson, Alexander Gutkin, Martin Jansche, and Linne Ha. 2018. Building Open Javanese and Sundanese Corpora for Multilingual Text-to-Speech. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Shijie Wu and Mark Dredze. 2019. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. In *2019 Conference on Empirical Methods in Natural Language Processing*. pages 833–844.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2019. BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis. *arXiv preprint arXiv:1904.02232*.

Keiji Yasuda and Eiichiro Sumita. 2008. Method for Building Sentence-Aligned Corpus from Wikipedia. In *2008 AAAI Workshop on Wikipedia and Artificial Intelligence (WikiAI08)*. pages 263–268.

Kuat Yessenov and Saša Misailovic. 2009. Sentiment analysis of movie review comments. In *Methodology*, 17: 1–7.

Bing Zhao and S. Vogel. 2002. Adaptive parallel sentences mining from web bilingual news collection. In *Proceedings of 2002 IEEE International Conference on Data Mining, 2002*. pages 745–748.