Proceedings of the 13th Workshop on Building and Using Comparable Corpora, pages 35–43
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Benchmarking Multidomain English-Indonesian Machine Translation

Tri Wahyu Guntara[1,2], **Alham Fikri Aji**[1,3], **Radityo Eko Prasojo**[1,4]
[1]Kata.ai, Jl. Kemang Raya No. 54, Jakarta, Indonesia
[2]University of Indonesia, Kampus UI Depok, Indonesia
[3]University of Edinburgh, Scotland
[4]Free University of Bolzano, Piazza Domenicani 3, {guntara, aji, ridho}@kata.ai

Abstract

In the context of Machine Translation (MT) from-and-to English, Bahasa Indonesia has been considered a low-resource language, and therefore applying Neural Machine Translation (NMT) which typically requires large training dataset proves to be problematic. In this paper, we show otherwise by collecting large, publicly-available datasets from the Web, which we split into several domains: news, religion, general, and conversation, to train and benchmark some variants of transformer-based NMT models across the domains. We show using BLEU that our models perform well across them and perform comparably with Google Translate. Our datasets (with the standard split for training, validation, and testing), code, and models are available on https://github.com/gunnxx/indonesian-mt-data.

Keywords: Neural machine translation, parallel corpus, English-Indonesian, Indonesian

1. Introduction

With approximately 200 million active speakers, Indonesian (*Bahasa Indonesia*) is the 10th most spoken language in the world (Eberhard et al., 2019). Yet, it is still considered to be one of the under-developed languages. Research in Indonesian Natural Language Processing (NLP) in general has suffered from a lack of open data, standardized benchmark, and reproducible code. Recent work in English-Indonesian (En-Id) machine translation (MT), in particular, has either used (1) closed data (Shahih and Purwarianti, 2016; Octoviani et al., 2019) or (2) open data with unpublished split for training, validation, and testing (Hermanto et al., 2015). Also, mostly only rule-based approaches or Statistical Machine Translation (SMT) were applied (Shahih and Purwarianti, 2016; Octoviani et al., 2019), whereas newer techniques such as Neural Machine Translation (NMT) based on the state-of-the-art Transformer architecture (Vaswani et al., 2017), which has been shown to outperform previous architectures such as the Recurrent Neural Network (RNN) in terms of training time and translation accuracy, has not been utilized. Hermanto et al. (2015) trained an RNN En-Id translation model. However, their model was trained only on a small amount of data with less than 24,000 parallel sentences. Furthermore, all these approaches have been evaluated using different datasets, and so it is unclear how well they perform in comparison to each other.

With the rise of the data-hungry NMT, effort such as the OPUS data portal (Tiedemann, 2012), OpenSubtitles (Lison et al., 2018), and Wikimatrix (Schwenk et al., 2019), has been made to publish more and more parallel data, including English-Indonesian to the number of millions of pairs. However, to the best of our knowledge, there has been no published work that utilizes the data for English-Indonesian machine translation. Therefore, in this particular context, it is currently unclear how useful the data is.

Bahasa Indonesia is a standardized register of Malay and is adopted as the country's national language to unify the archipelago with more than 700 indigenous local languages (Riza, 2008). Consequently, the daily-spoken colloquial Indonesian is vastly different from the standardized form due to the influences of the local language and, additionally, some popular foreign languages, such as English or Arabic. This phenomenon affects certain domains, such as the conversational domain where the colloquial Indonesian is typically used more, or the religion domain where Arabic words or phrases are sometimes used "as is" instead of being translated. Recent En-Id MT approaches have not yet considered different domains in Bahasa Indonesia (Shahih and Purwarianti, 2016; Octoviani et al., 2019) and instead have focused more on the news domain, which mostly used the standardized Indonesian (Hermanto et al., 2015).

In this work, our goal is to address the above problems by proposing several contributions as follow:

1. We collect scattered English-Indonesian parallel data available on the Web and divide them into several domains: news, religion, general, and conversation.
2. We introduce new datasets for news and conversation domains by aligning parallel articles and video captions.
3. For each domain, we set a standard data split for training, development, and testing. We further analyze the quality and characteristics of each dataset and each domain.
4. We train several transformer-based NMT models. We perform cross-domain testing to gain some insight into model robustness under domain changes. We conduct a manual evaluation of a sample of our data to assess the relative quality of our translation models further. We compare our results with Google Translate as the state-of-the-art translation tool.

The rest of the paper is structured as follow: Section 2 discusses the related work, which consists of parallel corpus collection and some En-Id MT approaches. Section 3 discusses the datasets that we use for training and testing. Section 4 describes the state-of-the-art and baseline MT methods that we use in our benchmark. Section 5 details our experiment settings and results, as well as discusses our findings and insights from the results. Finally, Section 6 concludes the paper and outlines some future work.

2. Related Work

The OPUS data portal (Tiedemann, 2012) provides a publicly available parallel dataset in 278 languages obtained from 55 open corpora,[1] although only 10 of them provide parallel data for English-Indonesian. Each corpus was collected from an open resource, and no manual data cleanup was carried out. Table 1 shows the statistics of the corpora containing English-Indonesian parallel sentences.

Corpus	doc's	sent's	en tok's	id tok's
OpenSubtitles v2018	9827	9.7M	72.8M	60.9M
Tanzil v1	45	0.5M	8.5M	15.4M
JW300 v1	8242	0.6M	10.0M	9.5M
Tatoeba v20190709	1	9.9K	11.0M	85.9K
QED v2.0a	2219	0.4M	4.8M	3.8M
GNOME v1	1347	0.5M	2.7M	2.3M
bible-uedin v1	2	62.2K	1.8M	1.4M
Ubuntu v14.10	398	96.5K	0.6M	0.3M
GlobalVoices v2017q3	562	14.5K	0.3M	0.3M
KDE4 v2	125	15.1K	86.0K	91.1K

Table 1: En-Id statistics shown on the OPUS webpage, November 2019

With over 9 million pairs, the OpenSubtitles dataset (Lison et al., 2018) represents around 80% of the En-Id sentence pairs in OPUS. The dataset is collected from the opensubtitles website.[2] Sentence pairs are extracted from two subtitles of different languages via time-slot alignment. Sometimes, there are time-slot mismatches because the subtitles are created using different sources of video with different play speeds and cut-off points. To combat the mismatches, two anchor points are selected as references to trim and to "stretch in/out" the other timestamps (Tiedemann, 2008).

Although OPUS is an open platform to publish parallel data, some dataset is not integrated in OPUS yet. Wikimatrix (Schwenk et al., 2019) collects 135 millions parallel sentences from Wikipedia across 85 languages. Multilingual sentence alignment of Wikipedia pages is done by leveraging LASER (Artetxe and Schwenk, 2019b), a massively multilingual sentence embeddings of 93 languages trained on a subset of OPUS. Using LASER, each sentence pair x and y of two different languages is scored using a margin formula that is a ratio of their cosine similarity and the average cosine of their k nearest neighbors, as follows:

$$\text{margin}(x,y) = \frac{\cos(x,y)}{\displaystyle\sum_{z\in \text{NN}_k(x)} \frac{\cos(x,z)}{2k} + \sum_{z\in \text{NN}_k(y)} \frac{\cos(y,z)}{2k}}$$

A margin threshold is applied to decide whether x and y are mutual translations or not. It has been shown to be more consistent than the standard cosine similarity in determining correct translation pairs (Artetxe and Schwenk, 2019a). Using this approach, Wikimatrix obtains at least 1 million En-Id sentences, depending on the threshold used.

Nevertheless, the data collected above has not yet been explored to build an English-Indonesian machine translation model. As English-Indonesian parallel data was considered to be low-resourced, attempts on data-driven machine translation are mostly a statistical-and-rule-based hybrid approach. Several examples include a general hybrid MT system where a rule-based morphological analysis is applied to generate an intermediate translation result which is then refined using an SMT model (Yulianti et al., 2011), a hybrid approach that analyzes Indonesian cliticization (Larasati, 2012a) and utterance disfluency (Shahih and Purwarianti, 2016) as a preprocessing step before feeding the training data into an SMT tool. Moving on from SMTs, Octoviani, et al. (2019) developed a neural-network-and-rule-based hybrid approach for phrase-based English-Indonesian Machine Translation. An RNN model is trained to classify the input phrase into a type. Then, a rule-based approach is applied for each phrase type to output the final translation. The approach was evaluated over a dataset of 70 pairs of phrases. Lastly, Hermanto et al.'s work (2015), which uses RNN, is the only work that we found within the topic of En-Id MT that utilizes NMT. They use the Pan Asia Networking Localization (PANL) dataset[3], which contains about 24,000 pairs of sentences, as their train and test data.

Due to the lack of distributed code from the previous work, we were not able to use them as our baselines. Instead, we use some variants of transformer-based models for our benchmark, which we will explain in details in Section 4.

3. Datasets

3.1. Existing Datasets

We collect data from OPUS (Tiedemann, 2012) which contains Open Subtitles (Lison et al., 2018) among other smaller datasets. Tanzil[4] and Bible-Uedin (Christodouloupoulos and Steedman, 2015) stores parallel Quran and Bible translations, respectively, while JW300 (Agić and Vulić, 2019) collects parallel sentences of Jehovah's Witness religious scripture and articles. Tatoeba[5] is a small database of sentences and translations in a general domain. GlobalVoices dataset[6] is a namesake of a multilingual news website,[7] from which its parallel sentences were crawled. Finally, GNOME[8], Ubuntu[9], and KDE4[10] datasets contain parallel software strings taken from their respective localization files.

We run the WikiMatrix (Schwenk et al., 2019) script to extract 1.8 million En-Id parallel sentences using a margin threshold value of 1.03 to obtain high-quality pairs in maximum number, as suggested in the paper. Other

[1] http://opus.nlpl.eu/ as of November 2019

[2] https://www.opensubtitles.org

[3] http://panl10n.net/english/OutputsIndonesia2.htm

[4] http://tanzil.net/trans/

[5] https://tatoeba.org/

[6] http://casmacat.eu/corpus/global-voices.html

[7] https://globalvoices.org/

[8] https://www.gnome.org/

[9] https://ubuntu.com/

[10] https://kde.org/

than OPUS and WikiMatrix, we find more, smaller datasets from the Web. The PANL dataset contains around 24,000 pairs of sentences manually aligned from news articles. IDENTIC (Larasati, 2012b) is a morphologically-enriched multidomain-dataset that combines the PANL dataset, a subset of Open Subtitles, and 164 manually-aligned sentences from BBC news articles. The Desmond86 dataset[11] contains parallel sentences obtained from BBC (news), Our Daily Bread (ODB)[12] (religion), SMERU[13] (research article), and AusAid[14] (humanitarian report). The Web Inventory of Transcribed and Translated Talks (WIT) (Cettolo et al., 2012)[15] released an extra dataset for the 2017 edition of International Workshop on Spoken Language Translation (IWSLT)[16], which also contains En-Id pairs extracted from TED talk videos. TALPCo contains high-quality pairs of short sentences originally translated from Japanese (Nomoto et al., 2018).

3.2. New Datasets

3.2.1. Bilingual BBC and BeritaJakarta

We use an earlier version of berita2bahasa.com crawler (Mitra, Sujiani and Negara, 2017) to crawl bilingual BBC[17] and bilingual BeritaJakarta[18] to extract parallel En-Id articles.[19] Each news article in the Bilingual BBC dataset is already paired and properly sentence-split. We observe that the translation style in this dataset is mostly one-to-one at the sentence level, meaning that most sentences are already paired. Although this results in less fluent translations in some cases, we have a straightforward sentence alignment with very few manual adjustments needed.

On the other hand, the Bilingual BeritaJakarta dataset is not yet aligned on the article-level. The Indonesian corpora contain 4000 timestamped articles, whereas the English contained 3000 articles. As the dataset was collected into a single clean text file, most of the article fingerprints are lost, and therefore using tools which rely on file fingerprints such as Bitextor (Esplá-Gomis and Forcada, 2009) is not feasible. We employ a timestamp-based alignment algorithm to find article pairs. First, for each language, articles published on the same date are grouped together. Then, two articles are paired following the order of publishing time, i.e., the first published article in Indonesian on a certain day is paired with the first published article in English on the same day, then the second article, then the third, etc. Mispairings are manually checked and fixed based on the titles. Then, we sentence-split the articles using NLTK (Loper and Bird, 2002). To ensure high-quality pairs, sentence alignment is performed manually.

3.2.2. Ibn Majah Parallel Translation

Sunan Ibn Majah is a major hadith[20] collection and has been translated into several languages. We crawled http://carihadis.com/[21] for the Indonesian translation and https://www.islamicfinder.org/[22] for the English one. However, the Indonesian source uses an older version of Ibn Majah, and therefore uses different hadith indexes, which makes an automated alignment problematic. Therefore, we perform manual alignment instead.

3.2.3. Youtube Parallel Caption

We extract YouTube videos whose captions are available in both English and Indonesian from several channels e.g., TED, TEDx, Khan Academy, Kobasolo, Raditya Dika, and Londokampung. Channels selected are based on our manual observation, that is, whether they contain a good portion of videos having both English and Indonesian captions. The Indonesian captions are transcribed directly, whereas the English captions are translated by their fans. A YouTube caption comes in a series of chunks where each chunk contains the text, the start time, and the duration of that particular chunk. The captions are not well-aligned since the length of parallel sentences in Indonesian and English differ, and only a small part of them can fit into the screen. But, unlike Open Subtitles, all pairs of captions on YouTube follow the same video source; thus, no timestamp stretch or cut-off is necessary.

Alignment is done using a greedy algorithm. First, chunks without timestamp intersection in the other language are discarded. Then, starting from the first pair of chunks, we compute how much time they overlap with each other. For instance, if an Id chunk starts from 0:00 and ends at 0:03, while an En chunk starts from 0:01 and ends at 0:04, then altogether they span 4 seconds but they occur at the same time for only 2 seconds. We say that they are together $2/4 = 50\%$ of the time. We call this measure as the intersection of union (IoU) ratio. We say that a pair of chunks are aligned if their IoU ratio falls above a certain threshold. If a pair of chunks do not satisfy the threshold, then the next chunk is appended to the shorter one among the pair, until the threshold is reached. We experimented with various threshold values on a small, randomly selected and manually annotated data, and found that 0.8 is a good threshold for aligning the chunks.

3.3. Dataset Analysis

We analyze the collected datasets for their quality and their domain characteristics. We quantitatively explore the datasets, as shown in Table 2. We mainly assess their quality based on their sentence lengths, unique tokens, noise, and completeness of sentences. We find that most of them are good quality. However, we find some other to be lacking, and decide to drop them. That is, they are not included in our benchmark.

[11] https://github.com/desmond86/Indonesian-English-Bilingual-Corpus. Sentence alignment was manually done, which was confirmed by the dataset owner via private messages.

[12] https://odb.org/

[13] https://www.smeru.or.id/

[14] defunct and now replaced by the Australian Aid

[15] https://wit3.fbk.eu/

[16] http://workshop2017.iwslt.org/

[17] https://www.bbc.com/indonesia/topik/dwibahasa, 2013

[18] beritajakarta.id, 2013

[19] https://herrysujaini.blogspot.com/2013/04/kumpulan-mono-korpus-bahasa-indonesia.html

[20] A kind of Islamic religious scriptures

[21] No ToC prohibiting crawling

[22] Content download is allowed for non-commercial uses

Corpus	**Abbr.**	$\lvert sent_{\text{en-id}}\rvert$	$\lvert tok_{\text{en}}\rvert$	$\lvert tok_{\text{id}}\rvert$	$\overline{len_{\text{en}}}$	$\overline{len_{\text{id}}}$	$\overline{len_{\text{ratio}}}$	**Domain/Content**
OpenSubtitles v2018	OpenSub	9.3M	0.4M	0.5M	7.72	6.41	1.32	Movie
*Tanzil v1	Tanzil	0.4M	24.3K	25.4K	21.47	33.05	2.06	Religion
JW300 v1	JW300	0.6M	87.6K	83.2K	17.44	16.26	1.20	Religion
*Tatoeba v20190709	Tatoeba	9.9K	5.7K	6.9K	7.63	6.62	1.23	General
QED v2.0a	QED	0.3M	82.8K	85.9K	14.65	12.95	1.33	Talk, Lecture
†GNOME v1	GNOME	40.4K	29.9K	30.1K	22.19	19.70	1.22	Tech
bible-uedin v1	Bible	59.4K	17.2K	21.0K	29.49	24.03	1.43	Religion
†Ubuntu v14.10	Ubuntu	96.5K	37.9K	44.2K	6.26	6.18	1.25	Tech
GlobalVoices v2017q3	GV	14.4K	27.5K	27.3K	21.06	18.94	1.21	News
†KDE4 v2	KDE	14.8K	9.5K	10.9K	5.72	6.26	1.49	Tech
Wikimatrix (T=1.02)	Wiki[x]	1.8M	1M	0.9M	22.75	21.06	1.22	General
∂Desmond86	Dsm	40.4K	29.9K	30.1K	22.19	19.7	1.22	News, Religion, Science
∂IDENTIC v1	IDENTIC	27.3K	36K	35.4K	22.96	21.29	1.20	News, Movie
IWSLT 2017	IWSLT	0.1M	48.7K	48.2K	19.67	16.85	1.23	Conversation
PAN Localization	PANL	24K	35K	35.5K	22.96	21.29	1.20	News
TALPCo	TALPCo	1.4K	1.2K	1.2K	9.08	7.58	1.26	General
BBC-BeritaJakarta	BBC-BJ	3.9K	10.5K	10.1K	20.36	18.36	1.22	News
†Ibn Majah	IbnMj	0.8K	3.9K	4.6K	65.41	51.95	1.4	Religion
YouTube v0	YT	0.3M	60.4K	63.4K	9.3	7.93	1.28	Talk, Lecture, Movie

Table 2: Exploratory data analysis of all datasets. Abbr. denotes the abbreviation of the corpus names. $|X|$ denotes the unique count of a set X, whereas $\overline{Y}$ denotes the average of bag of values Y. $\boldsymbol{len_{\text{ratio}}}$ denotes the absolute ratio between the sentence length of the two languages, En and Id. The absolute ratio between two arbitrary numbers x, y is $\max(x/y, y/x)$. Bold items indicate new datasets. †datasets that are dropped, ∂datasetes that are partially used, and *datasets with known problems but are used.

The Ubuntu and KDE4 datasets are taken from their respective software localization resources, and so we consider them to represent the tech domain. The majority of their "sentences" are short, incomplete, and noisy. For example:

- En: "%s: access ACL '%s': %s at entry %d"
- Id: "%s: akses ACL '%s': %s at masukan %d"

Therefore, the data as it is right now would not be very useful, and further refinement and filtering are necessary. The GNOME dataset, the third representative of the tech domain, unlike the other two, has higher-quality pairs. However, we could not find any other dataset within the same domain, so we decide to drop the tech domain altogether.[23] The Ibn Majah dataset contains sentences that are too long and need to be split, which is difficult due to inconsistent usage of splitting punctuations (commas, periods, colons, and semicolons) in the corpus. We decide to drop this dataset in our benchmark. The Desmond dataset contains a few numbers of pairs in the domain of Science, which are dropped. Lastly, the IDENTIC dataset has some intersection with the PANL and Open Subtitle datasets. Therefore we only consider the non-intersecting sentences.

After filtering out low-quality and redundant data, we combine the datasets falling under the same domain. News domain consists of news articles. Religious domain consists of religious manuscripts or articles. These articles are different from news as they are not in a formal, informative style. Instead, they are written to advocate and inspire religious values, often times citing biblical or quranic anecdotes. Next, we combine all datasets that come from human speech (movie, talk, and lecture) into the conversation domain. Lastly, we merge datasets that cover broad topics into the general domain. Then, for each domain, we split it into a train, validation, and test data. The result is shown in Table 3.

Domain	**Corpus**	**Sent's**	**Split**	$\mathbf{nsim_{V,T}}$
News	PANL	24k	train	3.3
	GV	14.4K	train	
	BBC-BJ	3.9K	valid+test	
Religion	Tanzil	0.4M	train	5.3
	JW300	0.6M	train	
	Bible	59.4K	train	
	Dsm_{ODB}	9k	valid+test	
Conversation	OpenSub	9.3M	all	18.5
	QED	0.4M	all	
	IWSLT	0.1M	all	
	YT	0.3M	all	
General	Wiki[x]	1.8M	all	7.3
	Tatoeba	9.9K	train	
	TALPCo	1.4K	all	

Table 3: Data split and n-gram similarity between validation and training data for each domain.

For news and religion domain, we choose an exclusive corpus:

- BBC-BJ for news, and
- Desmond ODB (Our Daily Bread, the religion part of Desmond dataset) for religion,

to be our validation and test data because (1) they are manually curated and of high-quality, (2) they are much smaller

[23] Experimentally, this is to avoid overfitting our model if it is trained on the tech domain with only one dataset.

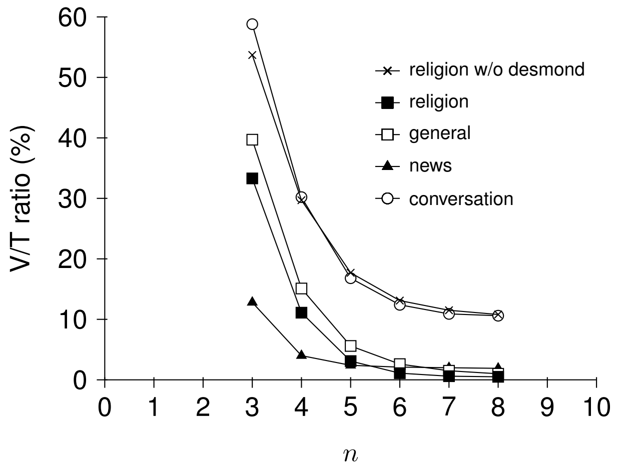

Figure 1: n-gram occurrences ratio between validation and test set across domains for n from 3 to 8.

than the rest of training data and therefore do not sacrifice too much portion of data that could have been for training instead, and (3) they have similar sentence length compared to the training data. There is no such corpus for the conversation domain and the general domain. The datasets in the conversation domain are all automatically aligned and therefore are noisy. For the general domain, both Tatoeba and TALPCo are manually curated, but their sentences (especially Tatoeba) are very short compared to Wikimatrix. Therefore, for these two datasets, we do a random split involving all datasets in the domain for validation and testing, each having 2000 unique pairs not present in the training set. For the general domain, we mix shorter sentences from TALPCo and the longer ones from Wikimatrix as our validation and test data. We observe that Tatoeba has similar types of high-quality sentences like TALPCo has, albeit shorter. Therefore we choose TALPCo to be in the validation and test sets instead, because longer sentences mean more difficult and meaningful evaluation.

To see the difference between these two split settings, we compute the rate of phrases (in terms of n-grams) that appear in validation set sentences that also appear in the training set sentences. Figure 1 shows this computation for $3 \leq n \leq 8$ for each domain. It shows that domains without an exclusive corpus for the validation set has a higher n-gram intersections between the validation set and the training set, which means that a model trained on the domain might be overfitted for the dataset and it might prove difficult to see how such a model generalizes to unseen dataset within the same domain. To further emphasize this point, we tried to built another split for the religion domain without the Desmond dataset, that is, the split involves all the other three datasets: Tanzil, Bible, and JW300. The result is that the validation and test sets share significantly more n-grams.

We further compute a weighted average of the occurrence ratios across ns, that is

$$\mathrm{nsim}(V,T) = \frac{\sum_{n=3}^{8} n \times 100 \frac{\mathrm{c}(n\mathrm{-gram\ in\ } V \mathrm{\ appearing\ in\ } T)}{\mathrm{c}(n\mathrm{-gram\ in\ } V)}}{\sum_{n=a}^{b} n}$$

where c is a counting function, V is the validation set, and T is the training set. The results of the weighted average of each domain is shown in Table 3, where the conversation domain is shown to have the highest $\mathrm{nsim}(V,T)$ of 18.5. In the next subsections, we discuss some special characteristics of each domain.

3.3.1. News

Some sentence pairs in the news domain suffer from the inter-sentence context-preservation issue. For instance, we sometimes find that a single sentence is aligned to two (usually shorter) sentences in the other language in order to capture the whole context of the single sentence. Another observation is the usage of pronouns, which loses context whenever the article is split into sentences and then paired. For example:

- En: **The firm** says the posts will go around ...
- Id: **Sony** mengatakan PHK karyawan dilakukan ...

In this example, ”Sony” as an entity is described as ”The firm”. Readers should understand the connection if presented with the whole article, but not as independent sentences.

Some sentences are appended with extra information to help the readers understand the news better based on their local knowledge. One of the most common examples is a converted currency, as shown in the example below.

- ”Kalau jauh misalnya di Indramayu, bisa 2,5 juta - 3 juta Rupiah.”
- ”If it is far, in Indramayu for instance, it could be around 2,5 - 3 million Rupiah ($**250 -** $**300**).”

Specifically, in Global Voices, we find translated tweets or Instagram posts, as this news site often include people’s reaction on social media in their articles. This part of the text is out-of-domain within the context of news. Furthermore, we find inconsistency in translating or copying the tweet’s usernames or tags.

3.3.2. Religion

The Tanzil dataset is a Quran translation dataset which has a relatively-imbalanced sentence length between the two languages, evidenced in Table 2, where an average Indonesian sentence in this dataset is about 50% longer than an average English one. Furthermore, an average pair of sentences in this dataset would, on average, have one of them twice as long as the other. However, we still decide to include the dataset in the domain to avoid overfitting because the remaining datasets are all about Christianity.

Another interesting property in the religion domain corpus is the localized names, for example, David to Daud, Mary to Maryam, Gabriel to Jibril, and more. In contrast, entity names are usually kept unchanged in other domains.

We also find quite a handful of Indonesian translations of JW300 are missing the end sentence dot (.), even though the end sentence dot is present in their English counterpart. Lastly, we also find some inconsistency in the transliteration, for example praying is sometimes written as ”salat” or ”shalat”, or repentance as ”tobat” or ”taubat”.

3.3.3. General

The Tatoeba dataset contains short sentences. However, they contain high-quality full-sentence pairs with precise translation and is widely used in previous work in other languages (Artetxe and Schwenk, 2019b). Due to its simplicity, we do not use Tatoeba as our test and validation sets. We find that the Wikipedia scraper for Wikimatrix is faulty in some cases, causing some noise coming from unfiltered markup tags.

3.3.4. Conversation

Our conversational domain corpus is translated from English. Hence the Indonesian sentences are written in formal language. In practice, Indonesian used informal language in speech, most of the time. In addition, we also used informal language in a conversational situation such as in social media or text messages.

4. Methods

4.1. Transformer-based Machine Translation

Transformer based model (Vaswani et al., 2017) is the current state-of-the-art for neural machine translation (Bojar et al., 2018). Therefore we adopt the standard Transformer-base encoder-decoder model as one of our baseline models.

4.2. Language-Model Pretraining

Generative pretraining has been proved to be effective in improving sentence encoders on downstream tasks. We use two language modeling objectives, Masked Language Modeling (MLM) to leverage our vastly available monolingual corpora and Translation Language Modeling (TLM) to make the network learns alignment between languages better. (Devlin et al., 2018; Radford et al., 2018; Lample and Conneau, 2019)

Although both MLM and TLM objectives can be extended to multiple languages, we only pretrain the base Transformer using Indonesian and English dataset since the network itself will only be used on tasks involving Indonesian and English languages. For the MLM objective, the Indonesian monolingual dataset was collected from Leipzig corpora (Goldhahn et al., 2012), and the English monolingual dataset was collected from WMT'07 and WMT'08.[24] Both datasets come from the news domain and are truncated at 4.8M sentences because of GPU resource limitation. For the TLM objective, Tatoeba and PANL datasets are used.

4.3. Google Translate

Google Translate is arguably one of the best public translation services available. However, benchmarking with Google Translate is tricky: Their model is regularly updated. Hence the result is not reproducible. We also cannot guarantee that our validation or test set is not present in their training data. However, we still argue that comparing our results with theirs is beneficial.

5. Experiments and Result

5.1. Setup

We run our Transformer experiment with XLM Toolkit on a single GPU. We use the Transformer base architecture, consisting of 6 encoder and decoder layers with 8 attention heads. The feed-forward unit-size is 2048, and the embedding size is 512. We increase the batch size from the default 32 to 160 to reduce the gradient noise (Wang et al., 2013; Smith et al., 2017), which shown to improve the model's quality (Ott et al., 2018; Popel and Bojar, 2018; Aji and Heafield, 2019). We use Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.0001, $\beta_1 = 0.9, \beta_2 = 0.999$. We train our language model with the same Toolkit. Performance is measured with a BLEU score (Papineni et al., 2002) by using sacreBLEU script (Post, 2018).

5.2. Model Evaluation

We first benchmark the significance of language-model pretraining for the Transformer. For this purpose, we train both vanilla Transformer and Transformer with language model pretraining for our news and general domain dataset. From the result shown in Table 4, we can see that the Transformer with language model pretraining outperforms its vanilla counterpart. We can also see that model trained in general domain outperforms model trained in news domain, therefore suggesting that a standard model with more data is better than a low-resource training with language model pretraining. For the next experiments, we will use a Transformer with a pretrained language model.

5.3. Cross Domain Evaluation

We explore the performance when trained across different domains. Our results shown in Table 5 suggest that the model is overfitted towards its specific domain. Model trained with the news domain dataset performed worst due to lack of resource. By combining every dataset, we can see the best performance across every domain. This result is comparable with Google Translate. We picked our best model, which is trained in all training set and evaluate the BLEU on test sets, which can be seen in Table 6.

5.4. Human Evaluation

We do not have an annotated parallel corpus for English-Indonesian. Our corpus, including the valid and test set, are generated from the crawled data. We discussed previously in section 3. that the currently available dataset are not fully parallel. Therefore, measuring the quality with BLEU only might not be representative.

For human evaluation, we select random sentences from each domain. We present three translations: Reference, Google Translate, and our output in random order to our human evaluators. We measure the quality in 2 scores:

- Fluency (1-5): How fluent the translation is, regardless of the correctness.
- Adequacy (1-5): How correct is the translation, given the source.

To ensure reliability of the scores, each and all sentences are assigned to 3 scorers. The final score is the averaged score across three evaluators, as shown in Table 7. Because we have more than two annotators and the scores are ordinal, we use Spearman's ρ to obtain a moderately-high average agreement between annotators of 0.53 for fluency and 0.56 for adequacy out of 240 sentences.

[24] http://www.casmacat.eu/corpus/news-commentary.html

Training Data	EN to ID evaluation (valid set)					ID to EN evaluation (valid set)				
	News	Religious	Conv	General	Average	News	Religious	Conv	General	Average
Transformer										
News	10.2	6.5	9.8	8.2	8.7	9.6	6.3	12.3	8.9	9.3
General	18.8	15.2	15.8	26.8	19.1	13.1	10.2	9.8	25.3	15.4
Transformer + Language Pretraining										
News	17.4	11.5	14.8	14.8	14.6	15.1	10.6	19.6	16.3	15.4
General	20.0	15.6	15.3	27.8	19.7	16.6	13.7	13.3	28.8	18.1

Table 4: Performance of different baselines across News (low-resource) and General (high-resource) domain.

Training Data	EN to ID evaluation (valid set)					ID to EN evaluation (valid set)				
	News	Religious	Conv	General	Average	News	Religious	Conv	General	Average
News	**17.4**	11.5	14.8	14.8	14.6	15.1	10.6	**19.6**	16.3	15.4
Religious	16.5	**21.5**	15.4	18.9	18.1	15.1	**20.2**	5.6	19.3	15.1
Conv	18.9	15.2	**28.0**	21.0	20.8	15.5	16.6	**33.1**	18.8	21.0
General	20.0	15.6	15.3	**27.8**	19.7	16.6	13.7	13.3	**28.8**	18.1

(a) Model generally performs well when evaluated with in-domain set. It performs poorly otherwise. An exception can be seen in the low-resource news domain.

Training Data	EN to ID evaluation (valid set)					ID to EN evaluation (valid set)				
	News	Religious	Conv	General	Average	News	Religious	Conv	General	Average
Transformer + Language Pretraining										
News + general	21.9	17.2	15.3	27.0	20.4	18.4	15.4	14.6	28.8	19.3
Relig.+ general	24.0	21.3	16.9	27.9	22.5	19.9	22.3	16.1	28.5	21.7
Conv + general	21.8	18.2	27.7	27.5	23.8	18.2	18.0	33.6	27.9	24.4
All	**24.6**	**21.6**	**27.8**	**28.1**	**25.5**	**20.5**	**22.5**	**33.3**	**27.9**	**26.1**
Google Translate										
-	25.0	23.8	27.0	26.3	25.5	25.0	29.1	28.9	28.8	28.0

(b) Adding general-domain to the training set improves the performance across different domains. Ultimately, combining all dataset yields the best results.

Table 5: Cross-domain evaluation of Transformer with language pretraining

Test Domain	EN to ID	ID to EN
News	24.4	20.2
Religious	21.3	22.1
Conversation	27.3	32.4
General	28.1	28.9
Average	25.3	25.9

Table 6: Evaluation on test set. We compare our model trained with all dataset with Google Translate (GT).

	News	Relig.	Conv	General	Avg
Fluency					
Corpus	4.78	4.73	4.63	4.63	4.69
Ours	4.44	4.22	4.62	4.21	4.37
Google	4.26	3.85	4.53	3.59	4.06
Adequecy					
Corpus	4.34	4.58	3.92	3.92	4.19
Ours	4.05	4.09	4.38	4.1	4.15
Google	4.27	3.99	4.6	3.92	4.2

Table 7: Human evaluation score across different domains.

The reference translation is the most fluent across every domain. This result is expected, as the reference is written by humans. Reference translation's adequacy scored equally on average, compared to the rest. Our reference is crawled; therefore, it contains several issues, as mentioned in section 3.3.. One main problem in reference translation is that they are translated with document level in mind, therefore reducing adequacy as encapsulated sentence-based translation. This is especially true in conversational, where the reference was translated from the whole session (i.e., talk, or vlog). One example can be seen below:

Source	”- Nope, they're shutting us down.”
Ref	”- Tidak, misi ditunda.”
Ours	”- Tidak, mereka menutup kita”.
Google Translate	”- Tidak, mereka menutup kita.”

The reference is literally translated as ”- No, mission postponed.”, which is not the correct translation of the source. However, the reference is in fact acceptable when given the whole document.

6. Conclusions and Future Work

We showed that Bahasa Indonesia has improved from the preconception of being a low-resource language in the context of English MT. We have collected scattered English-Indonesian parallel data and introduced some new parallel datasets through automatic and manual alignments. Our collected datasets numbers in more than 10 million pairs of sentences. We evaluated and categorized those datasets into several domains: news, religion, general, and conversation. We created a standardized split for evaluation to open a pathway for objective evaluation for future En-Id MT research. Our Transformer-based baseline trained with multidomain dataset produces a comparable quality compared to Google Translate and is robust against domain changes. However, we acknowledge that some improvements to our datasetes are necessary. Some important domains like news are still behind in terms of training data, and evidently, its BLEU score is still lacking compared to the general and conversational domain. Furthermore, our manual evaluation has shown that some of our datasets contain noise, especially in the conversation and general domain where the noisy data is still used in validation and testing. In the future, manual data filtering or cleansing on these datasets is important to ensure that we have a standard benchmark that is clean and unbiased.

7. Bibliographical References

Agić, Ž. and Vulić, I. (2019). Jw300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210.

Aji, A. F. and Heafield, K. (2019). Making asynchronous stochastic gradient descent work for transformers. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 80–89, Hong Kong, November. Association for Computational Linguistics.

Artetxe, M. and Schwenk, H. (2019a). Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy, July. Association for Computational Linguistics.

Artetxe, M. and Schwenk, H. (2019b). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., and Monz, C. (2018). Findings of the 2018 conference on machine translation (wmt18). In *Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers*, pages 272–307. Association for Computational Linguistics, 10.

Cettolo, M., Girardi, C., and Federico, M. (2012). Wit3: Web inventory of transcribed and translated talks. In *Conference of European Association for Machine Translation*, pages 261–268.

Christodouloupoulos, C. and Steedman, M. (2015). A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49(2):375–395.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eberhard, D. M., Simons, G. F., and Fennig, C. D. (2019). *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, 22nd edition.

Esplá-Gomis, M. and Forcada, M. L. (2009). Bitextor, a free/open-source software to harvest translation memories from multilingual websites. *Proceedings of MT Summit XII, Ottawa, Canada. Association for Machine Translation in the Americas*.

Goldhahn, D., Eckart, T., and Quasthoff, U. (2012). Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*, volume 29, pages 31–43.

Hermanto, A., Adji, T. B., and Setiawan, N. A. (2015). Recurrent neural network language model for english-indonesian machine translation: Experimental study. In *2015 International Conference on Science in Information Technology (ICSITech)*, pages 132–136. IEEE.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Larasati, S. D. (2012a). Handling indonesian clitics: A dataset comparison for an indonesian-english statistical machine translation system. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 146–152.

Larasati, S. D. (2012b). Identic corpus: Morphologically enriched indonesian-english parallel corpus. In *LREC*, pages 902–906.

Lison, P., Tiedemann, J., and Kouylekov, M. (2018). Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Loper, E. and Bird, S. (2002). Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.

Mitra, V., Sujaini, H., and Negara, A. B. P. (2017). Rancang bangun aplikasi web scraping untuk korpus paralel indonesia-inggris dengan metode html dom. *Jurnal Sistem dan Teknologi Informasi (JUSTIN)*, 5(1):36–41.

Nomoto, H., Okano, K., Moeljadi, D., and Sawada, H. (2018). Tufs asian language parallel corpus (talpco). In *Proceedings of the Twenty-Fourth Annual Meeting of the Association for Natural Language Processing*, pages 436–439.

Octoviani, W., Fachrurrozi, M., Yusliani, N., Febriady, M., and Firdaus, A. (2019). English–indonesian phrase translation using recurrent neural network and adj technique. In *Journal of Physics: Conference Series*, volume 1196, page 012007. IOP Publishing.

Ott, M., Edunov, S., Grangier, D., and Auli, M. (2018). Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Popel, M. and Bojar, O. (2018). Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70.

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

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Riza, H. (2008). Resources report on languages of indonesia. In *Proceedings of the 6th Workshop on Asian Language Resources*.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2019). Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Shahih, K. M. and Purwarianti, A. (2016). Utterance disfluency handling in indonesian-english machine translation. In *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, pages 1–5. IEEE.

Smith, S. L., Kindermans, P.-J., Ying, C., and Le, Q. V. (2017). Don't decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489*.

Tiedemann, J. (2008). Synchronizing translated movie subtitles. In *LREC*.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Wang, C., Chen, X., Smola, A. J., and Xing, E. P. (2013). Variance reduction for stochastic gradient optimization. In *Advances in Neural Information Processing Systems*, pages 181–189.

Yulianti, E., Budi, I., Hidayanto, A. N., Manurung, H. M., and Adriani, M. (2011). Developing indonesian-english hybrid machine translation system. In *2011 International Conference on Advanced Computer Science and Information Systems*, pages 265–270. IEEE.